AFCAPS-FR-2011-0009

A SUMMARY OF THE TECHNICAL PILOT SELECTION LITERATURE

Diane L. Damos
Damos Aviation Services, Inc.

Sponsored by
HQ AFPC/DSYX & HQ AF/A1PF
Mr. Kenneth L. Schwartz
Strategic Research and Assessment Branch

October, 2011

Air Force Personnel Center
Strategic Research and Assessment
HQ AFPC/DSYX
550 C Street West, Ste 45
Randolph AFB TX 78150-4747

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report was cleared for release by HQ AFPC/DSYX Strategic Research and Assessment Branch and is releasable to the Defense Technical Information Center. The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report is published as received with minor grammatical corrections. The views expressed are those of the authors and not necessarily those of the United States Government, the United States Department of Defense, or the United States Air Force. In the interest of expediting publication of impartial statistical analysis of Air Force tests SRAB does not edit nor revise Contractor assessments appropriate to the private sector which do not apply within military context.
Federal Government agencies and their contractors registered with Defense Technical Information Center should direct request for copies of this report to:

Defense Technical Information Center - http://www.dtic.mil/

Available for public release. Distribution Unlimited. Please contact AFPC/DSYX Strategic Research and Assessment with any questions or concerns with the report.

This paper has been reviewed by the Air Force Center for Applied Personnel Studies (AFCAPS) and is approved for publication. AFCAPS members include: Senior Editor Dr. Thomas Carretta AFMC 711 HPW/RHCI, Associate Editor Dr. Gregory Manley HQ AFPC/DSYX, Dr. Lisa Hughes AF/A1PF, Dr. Paul DiTullio AF/A1PF, Kenneth Schwartz HQ AFPC/DSYX, Johnny Weissmuller HQ AFPC/DSYX, Dr. Laura Barron HQ AFPC/DSYX, Dr. Mark Rose HQ AFPC/DSYX, and Brian Chasse HQ AFPC/DSYX.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
17-10-2011	Technical	August 2009-October 2011

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Summary of the Technical Pilot Selection Literature	W911NF-07-D-0001
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S).	5d. PROJECT NUMBER
Damos, Diane L.	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT
Damos Aviation Services, Inc. 36303 N. Old Woods Trail Gurnee, IL 60031	DAS-2011-05
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
Air Force Personnel Center **Strategic Research and Assessment Branch** **Randolph AFB TX 78150**	HQ AFPC/DSYX
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
	AFCAPS-FR-2011-0009

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release, Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This report summarizes the history of US military pilot selection for fixed-wing aircraft beginning in 1917 with the entry of the United States into World War I. The World War I effort to develop a predictive pilot selection battery is covered in detail. The second chapter of the report deals with the period between the end of World War I and the beginning of World War II. The little research that was conducted during this period is presented along with changes in the Army's and Navy's pilot selection batteries and processes. The impact of a reduced selection system on training outcomes is described. The third chapter deals with pilot selection in World War II and describes the development of the (cont.)

15. SUBJECT TERMS
aviator selection, KSAOs, personality, psychomotor, ability, history of psychology, biodata, classification, screening

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
Unclassified					Kenneth L. Schwartz
a. REPORT	b. ABSTRACT	c. THIS PAGE		54	19b. TELEPHONE NUMBER *(include area code)*
U	**U**	**U**	**U**		210-565-3139

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

14. Abstract (continued)
Army Air Forces' pilot selection system and the Navy's system. Because of the large amount of material available on pilot selection during this period, this chapter summarizes the research and development efforts and presents representative predictive validities for the selection instruments. The fourth chapter deals with the 5-year period between the end of World War II and the beginning of the Korean War. Few developments occurred during this period, and the chapter is correspondingly brief. The final chapter presents an overview of work conducted since 1950. It concentrates on three types of selection instruments that have been used since World War II: personality, biographical, and timesharing. This chapter ends with a final perspective on the efforts to develop a predictive pilot selection system and suggests areas for further research.

ACKNOWLEDGEMENT

This work was supported by the Air Force Personnel Center (AFPC/DSYX; Brian G. Chasse) under the auspices of the U.S. Army Research Office Scientific Services Program administered by Battelle (Delivery Order 0762, Contract No. W911NF-07-D-0001). The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

TABLE OF CONTENTS

TABLES

ABSTRACT

This report presents the history of US military pilot selection for fixed-wing aircraft beginning in 1917 with the entry of the United States into World War I. The World War I effort to develop a predictive pilot selection battery is covered in detail. The second chapter of the book deals with the period between the end of World War I and the beginning of World War II. The little research that was conducted during this period is presented along with changes in the Army's and Navy's pilot selection batteries and processes. The impact of a reduced selection system on training outcomes is described. The third chapter deals with pilot selection in World War II and is divided into a section describing the development of the Army Air Forces' pilot selection system and a section dealing with the Navy's system. Because of the large amount of material available on pilot selection during this period, this chapter summarizes the research and development efforts and presents representative predictive validities for the selection instruments. The fourth chapter deals with the 5-year period between the end of World War II and the beginning of the Korean War. Few developments occurred during this period, and the chapter is correspondingly brief. The final chapter presents an overview of work conducted since 1950. It concentrates on three types of selection instruments that have been used since World War II: personality, biographical, and timesharing. This chapter ends with a final perspective on the efforts to develop a predictive pilot selection system and suggests areas for further research.

A SUMMARY OF THE TECHNICAL PILOT SELECTION LITERATURE

EARLY MANNED FLIGHT

The first manned flight was performed by the Montgolfier brothers on June 4, 1783. The balloon was tethered and used hot air for buoyancy. This ascent was followed quickly by the first free flight on November 21, 1783 and by the first free flight of a hydrogen-filled balloon on December 1, 1783. All three flights occurred in France. The first balloon flight in the United States occurred January 10, 1793 and was observed by George Washington. The balloonist, however, was French, not American.

No information was located on the use of balloons in the United States until the Civil War, when balloons were used for military purposes. During the Civil War balloons were tethered and used for observation. The only information about selection for balloonists at this time is found in the memoirs from a Confederate balloonist (Farr, 1993) who flew in April, 1862. The selection process consisted of an interview with the commanding general, who was concerned only that the observer was familiar with the local terrain and could distinguish one branch of the Army from another. The military continued its interest in balloons through World War I (see for example *Annual Report of the Secretary of War for the Year 1885. Vol III: Report of the Chief of Ordnance*, 1885). However, no information was located on selection of balloonists between the end of the Civil War and the beginning of World War I.

The earliest pilot selection efforts for fixed-wing aircraft were concerned with medical issues associated with flying. The first paper on aviation-related medical issues appeared in 1907 (Anderson, 1919), only 4 years after the Wright brothers' first flight. In 1911, a U.S. Army physician made recommendations for a special physical examination for aviators (Mashburn, 1939), which resulted in an aviator physical for the U.S. Army in 1912. During this early period, the medical community recognized that psychiatric problems could affect flying. In 1914, Ovington, an American, published a letter in the *Journal of the American Medical Association* entitled "The Psychic Factor in Aviation," although the letter did not actually address any psychiatric or psychological factors that might affect flying performance. By 1916, the Royal Air Corps had documented nervous breakdowns of cadets (Anderson, 1919) but did not screen applicants for psychiatric problems. The development of a pilot selection system that included physical, psychiatric, biographical, and cognitive factors had to await the beginning of World War I.

WORLD WAR I

Background

Modern psychologists understand how to develop a selection process and identify the most efficacious selection instruments. Prior to and during World War I, psychologists did not understand the process for developing a selection system and had little experience with selection instruments. Consequently, the development of the U.S. pilot selection system from its inception through the end of World War I followed a path that is unlike that of a modern selection system. Many steps in the modern process were not present and many pitfalls were unrecognized at the time. Additionally, the initial development of the U.S. military pilot selection system faced some unique problems. The most prominent of these were the complete lack of a scientific establishment with any pertinent expertise and the absence of a pilot selection database.

To understand the development of the pilot selection system, one must first understand the development of the U.S. Army selection system for recruits; the pilot selection system was embedded in the larger Army system. The broad outlines of the Army system development are detailed in two articles by Yerkes (1918, 1919) and will be described in detail later. Unfortunately, Yerkes' articles provide little information on the development of the pilot selection system. Although the outline of the process is clear, some of the details are murky and some of the dates have to be inferred from conflicting sources. Nevertheless, the major problems in understanding the development of the system lie with the selection instruments. The sources of the tests and the rationale for their inclusion in the selection battery are often unclear. To muddy the water further, the psychologists involved in test development for the Army used different names for the same test. Finally, many of the documents pertaining to the development of the selection system do not give exact dates for the validation trials or for the development of new tests. Thus, the timeline for test development and implementation had to be reconstructed from several different sources.

The second set of problems concerns terminology. During the period just before and during World War I, many cognitive and psychomotor skills and abilities were conflated with physiological factors. The terminology used during this time differed substantially from modern usage, making it difficult to determine exactly what was being tested.

The following sections will present the development of the Army selection system first and then the development of the pilot selection system.

Development of the Psychological Selection System for U.S. Army Recruits

In the afternoon of April 6, 1917, the day the United States declared war on Germany, a group of psychologists know as the "experimentalists" was holding a meeting in Cambridge, MA. (Yerkes, 1918). The next day, this group drafted a letter to the American Psychological Association (APA) asking the board of the Association to appoint a committee to determine how psychology could assist the war effort. To identify possible areas where psychology could assist, Robert M. Yerkes (who was both a member of the experimentalists and president of APA) traveled to Canada on April 10th. He was particularly interested in selection for Army recruits and rehabilitation of wounded soldiers. Yerkes asked Canadian military officers about psychological selection for pilots and was told that Canada had no "significant" psychological tests in their selection battery (p. 90).

On April 22nd, the APA board voted to have Yerkes establish 12 committees and appoint a chair for each. Each committee was concerned with a specific area of psychology that could contribute to the war effort and was instructed to render all possible aid to the U.S. government. Two committees are important for pilot selection. The first was chaired by Yerkes on the "Psychological Examination of Recruits." The second was on "Psychological Problems of Aviation, including Examination of Aviation Recruits." This committee initially was chaired by H.E. Burtt, followed by G. Stratton, and finally by E.L. Thorndike (Yerkes, 1919). This committee also included John B. Watson, Warner Brown, Francis Maxfield, and H.C. McComas. The most prominent committee members were university professors with laboratories and active research programs. Several of these individuals appear to have given the military tests they had developed. This transfer, with rare exceptions, was not documented and must be inferred by comparing the descriptions of the tests in the pilot selection battery to prior publications by these individuals. The April 22nd meeting also produced a memo to all APA members urging them to offer their services and all of the resources of their laboratories to the government.

At this meeting the board voted to place its committees under the National Research Council (NRC). Subsequently, the NRC established a psychological committee in conjunction with the APA, the National Academy of Science, and the American Association for the Advancement of Science (Yerkes, 1918). This committee was to supervise psychological research done for the war effort. Issues proposed by military officers or other psychologists were directed to the appropriate institution or individual for immediate attention.

By the last week of April, 1917, Yerkes had developed a plan for the psychological evaluation of recruits. In May 1917 Yerkes' plan was sent to the Surgeon General of the Army. The chair of the Committee of Medicine and Hygiene of the National Research Council felt so strongly about selection of army recruits that he decided not to wait for government funding for selection research. Instead, he obtained private funding for Yerkes to develop selection tests. Yerkes subsequently assembled a team that began work on May 28, 1917. By July 7th, the team had developed and revised both group and individual examinations. On December 24, 1917 the Army accepted the concept of psychological testing for all recruits and began to develop a plan for such testing.

Two of the purposes of the psychological testing were to identify men for officer training and for special assignments. Although Yerkes never specifically mentioned selection for any flying-related duties, by the end of the war, commanders of flying training fields in Europe allowed enlisted men serving at the training fields to begin flight training. These individuals had been promised flight training as a result of their test results.

First Selection System—Form 609

The U.S. Army Air Services revised its 1912 medical examination under the direction of Cols. T.C. Lyster and Isaac Jones in May, 1917 (McFarland, 1953, p. 88). This revised version became known by the form used to record data, "Form 609." The use of Form 609 was a major advance in pilot selection; the results of all flight examinations, regardless of where they were conducted, were recorded on the same form. The adoption of this form was accompanied by an equally revolutionary process—the development of standardized examination centers. Three American

physicians convinced the Army in early 1917 to distribute these centers across the country, and one of these three, Dr. Isaac Jones, began to establish the Midwest centers in July 1917 (Jones, 2008). Where possible, these centers were placed in large hospitals or state universities. The medical equipment was standardized and the same examination techniques were used in each location. The examination was based on the concept of an assembly line with a fixed amount of time allocated for each candidate. Local volunteer physicians staffed the centers, with a medical officer in charge. Each of these units examined between 10 and 60 applicants each day. Eventually, medical units for the selection of pilot applicants were established in 35 cities across the United States (Anon., 1919). A second set of 32 medical units was established at Army training camps to administer the tests to enlisted applicants (Anon., 1919). The success of Form 609 in supplying a sufficient number of candidates for flight school is evinced in the fact that the Army stopped all pilot applicant examinations on February 9, 1918 because the number of successful pilot applicants waiting to begin training greatly exceeded the capacity of the training facilities.

By the end of World War I the selection process associated with Form 609 included a physical examination, a psychiatric examination, and a variety of non-medical selection instruments. The administration and scoring of the medical and psychiatric examinations was under the control of the medical officer of each facility. The non-medical instruments were administered and scored by Aviation Examining Boards. As will be evident in the discussion below, non-medical selection instruments were added throughout the war. Unfortunately, the point at which each of the instruments was added cannot be determined with certainty.

Until July, 1918 all candidates for flight training were officers. Consequently, no applicant could be accepted for flight training if he had not completed high school or preparatory school and could provide the necessary documentation. If the candidate had a college degree or the equivalent, the non-medical portion of Form 609 could be omitted. Johnson (1917) pointed out that a college degree did not guarantee the types of abilities required by flying and strongly suggested that the correct combination of psychological tests would increase the probability of identifying successful candidates. His suggestion about requiring all candidates to complete the non-medical portion of Form 609 appears to have gone unheeded.

Physical examination. The medical examination was comprehensive and not altered until after World War I (Anon., 1919). It consisted of a medical history, an examination of the circulatory system, a urine analysis, and a basic blood panel. The applicant's vision and hearing also were tested and his teeth examined.

The U.S. Army clearly understood that careful medical selection of applicants decreased training costs (Anon., 1919). Consequently, the physical standards were rigorous. Armstrong (1943) indicates that 30.3% of the pilot applicants failed one or more parts of the medical examination.

Psychiatric examination. The selection process appears to have included a psychiatric interview from its adoption. Stratton, McComas, Coover, and Bagby (1920) indicate that the psychiatric examination was distinct from all of the other "mental tests" that were subsequently included in the selection process. No other descriptions of the psychiatric examination were found, but two versions of Form 609 have been located. One was modified for use by the

American Expeditionary Force in Europe (Wilmer & Ireland, 1920). This version was "improved," and may have been used only to qualify enlisted men who had been promised flight training on the basis of their prior test scores (Army Alpha plus others). This version asked only about prior "mental or nervous breakdowns," and the candidate's and his family's attitude about choosing the Air Service. Only one line was available for describing the candidate's personality. The second was included in an immediate post-war Air Service medical manual (Anon., 1919). This version had no psychiatric or personality questions.

Non-medical selection instruments. The few descriptions of the non-medical portions of the selection process are particularly vague, brief, and conflicting (Henmon, 1919; McComas, 1922; Stratton et al., 1920). Henmon (1919) implies that the selection process always contained non-medical elements and refers to "mental, moral, and professional requirements (p.103)." In contrast, McComas (1922) mentions only medical tests in Form 609 and implies that other tests were administered after the student reported for preliminary flight training.

These differences may be attributed to rapid changes in the non-medical portion of the selection process and to changes in the validation process. During the early phase of the war, huge number of individuals volunteered and underwent ground school and flight training. Later, the selection process was halted because of a backlog of trainees. All subsequent validation had to be conducted at ground schools and flight schools.

As noted earlier, 30.3% of the pilot applicants failed one or more parts of the medical examination (Armstrong, 1943). Henmon (1919) estimates that between 50 and 60% of the pilot applicants failed the selection process (either the medical or the non-medical portions). These figures imply that at least 30% of the candidates who passed the medical portion of Form 609 failed the non-medical portion.

Professional, mental, and moral examinations. Most sources (Henmon, 1919; McComas, 1922; Stratton et al, 1920) mention mental, professional, and moral examinations and give the impression that these three examinations were always included in Form 609. Stratton et al. (1920) provides a vague description of the "professional and mental examination" and indicate that the mental examination was distinct from the psychiatric examination included in Form 609 and from the "special tests of mental alertness," which will be discussed later. The mental and professional examinations included "the candidate's written answers to a series of questions covering his parentage, education, business experience, athletic attainments, and responsibilities placed upon him by others, military training….(p.406)" As part of these examinations, the candidate had to submit to an oral examination by the Aviation Examining Board.

No documents were located describing either the professional, mental, or oral examinations in detail or providing sample questions. The Form 609 used by the American Expeditionary Force in Europe includes a single or a half-line answer space for questions pertaining to the applicant's occupation and participation in sports. Although the Form 609 used in Europe was not the same as the form used in the United States, the short answer spaces combined with the lack of documentation suggest that the professional and mental examinations were unstructured interviews that focused on these short answers. The major argument against a completely

unstructured interview rests on John B. Watson, G.M. Stratton, and V.A.C. Henmon. Watson enlisted in the Army in 1917 and was put in charge of the "organization of methods" (Yerkes, 1918) of the non-medical portion of the Aviation Examining Board. Both Stratton and Henmon, who would play major roles in the development of the first pilot selection battery, served on these boards. The need for reliable information to guide the decisions of the Examining Boards was realized by June, 1917. Thus, the assumption that no attempt was made to standardize the mental and professional examination seems tenuous but must await further research.

The moral examination apparently consisted of a careful review by the Board of at least three letters of recommendation for a candidate attesting to his moral character and fitness for duty in the Air Service. These letters of recommendation were to be written by individuals with a detailed knowledge of the candidate.

Mental Alertness Test. One of the most promising pilot selection tests of World War I was the Mental Alertness Test, which was developed by E.L. Thorndike. Thorndike (1919) realized that success as a military pilot required that the candidate 1) pass ground school, 2) pass flight training, 3) possess the personality traits and background of an officer, and 4) function as a military aviator at the front. He studied how the Aviation Examining Boards identified good pilot candidates and noted that years of education was the best predictor of success in ground school. He subsequently developed a "systematic test of intelligence or mental alertness," which was composed of 13 subtests (Henmon, 1919). Unfortunately, no description of these 13 tests was found.

According to Thorndike the Mental Alertness Test correlated .50 with success in ground school, whereas years of education correlated approximately .25 with success in ground school (note that "success" is not defined). The correlation between success in ground school and score on the Mental Alertness Test with years of education partialled out was approximately .4. Thorndike then determined that the test scores correlated .3 with "ability to learn to fly" and .3 with "general officer-quality," neither of which is defined. Thorndike provides no information on these correlations, which may be based on a composite of the 13 subtests or on one specific subtest.

Thorndike (1919) also does not mention when he developed the Mental Alertness Test or when it was adopted by the Aviation Examining Boards. Henmon (1919) provides a few vague and confusing comments about the adoption of the Mental Alertness Test that point to its incorporation into the selection process prior to March, 1918. In contrast, Yerkes (1919) states that Thorndike began work on the Mental Alertness Test in August, 1918. Given that Thorndike obtained correlations with both ground school and with flight performance, this later date seems unlikely unless the test was developed in a few weeks and administered immediately both to cadets at the beginning of ground school and to cadets at the beginning of preliminary flight training. In such a scenario, the correlation between test score and flying performance would be an underestimation caused by range restriction; only those cadets who successfully completed ground school took the test.

Thorndike-Kelly Athletic Interest Inventory. Thorndike also developed and validated a test of athletic interests and achievement (Henmon, 1919). The test was added to the non-

medical portion of Form 609 sometime between August, 1917 and March, 1918. Henmon provides the only comment about the development process, "Thorndike's study of application blanks filled out by candidates and subsequent performance had shown a positive correlation between athletic ability and success in flying. A more detailed blank was prepared and a scoring system for it worked out by Thorndike and Kelly (p. 106)." Henmon also reports that it correlated .6 with the ratings but gives no source for the data.

Thorndike (1919) provides no description of the test or the items, and again no example of this test was located. This lack is particularly disappointing because the Athletic Interest Inventory was a foundational test. Variations of some of the test items are still in use today by the U.S. Army for pilot selection (HQDA, 1987).

Developing Apparatus-based Pilot Selection Tests—Spring, 1917-Spring, 1918

Although 60% of the early pilot candidates failed one or more parts of the selection process, the attrition rate in flight training was higher than the Army could tolerate. Members of the APA/NRC committees quickly recognized that additional pilot selection tests were needed to decrease the attrition in training and assembled a battery of 23 apparatus tests (tests that required either some type of apparatus for stimulus presentation or for recording the candidate's response). This battery was administered to cadets at the ground school at the Massachusetts Institute of Technology (MIT) in early June, 1917 (Yerkes, 1919). Yerkes provides a one-sentence description of some of the tests but never discusses how these tests were chosen; he only states that the tests "promised *a priori* to be indicative of aptitude for flying (p. 95)." Because derivatives of these tests are still in use today, it is important to identify their source and the constructs measured by the instrument.

Henmon (1919) indicates that some of the tests were considered to be promising by French and Italian psychologists. However, only three tests are clearly derived from European selection batteries (See Appendix A for a description of European selection) and the immediate source of these tests is unclear; direct contact between American and Allied psychologists is documented only during the fall of 1918 when a group of American uniformed psychologists was sent to air fields in France. One test assessed simple reaction time to visual stimuli; the second, to auditory stimuli. The third test, which was assumed to assess "emotional stability," recorded changes in the applicant's pulse rate, breathing rate, and "arithmetic performance" after a revolver was fired out of sight of the applicant. No information is given about how these changes were scored.

At least 12 of the 20 remaining tests appear to measure medical or physical factors—patellar reflex, visual acuity, auditory threshold, a primitive Galvanic skin response, cardiograms taken during exercise, postural sway, equilibrium (response to tilt), time to muscle fatigue, and four measures of other visual functions. Descriptions of two other tests involving simple finger movements are inadequate to determine if the tests assess physical or basic psychological factors. The source and purpose of these 14 tests is unclear, especially given that the medical portion of Form 609 was developed by this point. One possibility is the members of the APA/NRC committee selected these tests because they were familiar with them; several members were, in today's terminology, biological psychologists or sensation and perception psychologists. The visual tests appear to have been derived from previous work by some of these members, but no direct link was located.

Two tests in the battery may have assessed spatial abilities. One, the distance and velocity estimation test, appears to have assessed dynamic spatial reasoning using the same method employed today: The applicant observed an object moving at a constant rate for a short period of time. The object then disappeared, and the applicant had to indicate when the object would reach a given point. The second test, a maze learning test, may have assessed visualization. This test required the applicant to trace a maze. The applicant then re-traced the maze when the maze was not visible and then traced it when it was not visible and had been rotated. No sources for these two tests were identified.

Two other tests assessed verbal abilities. One test appeared to be a type of verbal memory test; words were exposed one letter at a time. The description does not indicate what the applicant did or what response was measured. Another test, the "association reaction with crucial words" may be a variation of the association reaction task developed by Knight Dunlap as a test of concept learning. Dunlap, a psychology professor, apparently developed this test to determine how well the students in his courses had learned certain concepts. His 1917 paper indicates that a student was given a word and had to respond with another, related word. His response was graded on similarity to the stimulus word. Dunlap was a member of one of the APA/NRC committees prior to volunteering and worked in the Army laboratory that conducted the majority of work on pilot selection after enlisting. How this test was modified for use in pilot selection is unclear, as is its rationale for inclusion in the battery.

One of the final two tests was a choice reaction time task. This was described as a "continuous" choice reaction time task. Apparently, a new stimulus was presented as soon as the candidate made a response. The last test was a two-alternative "motor learning" test, in which the candidate learned a sequence of responses by trial and error. The dependent measure is not given. Based on the one sentence description, the assumption that this test measures motor learning may be questioned.

The APA/NRC Committee on "Psychological Problems of Aviation, including Examination of Aviation Recruits" administered the battery to 75 cadets at MIT, but flying training records were obtained only for 25. Four more data collection efforts were conducted. In August, 1917, Stratton tested over 50 cadets in San Diego (Yerkes, 1919; Henmon, 1919) using a subset of the MIT battery. He used simple reaction to auditory and visual stimuli, the "emotional stability" test described above, the postural sway test, and the tilt test. He also used a "dexterity" test, which may have been one of the tests in the MIT battery. Additionally, he appears to have used the test of dynamic spatial reasoning. However, Stratton's test stimuli are described as moving in curved paths, whereas the stimuli for the MIT battery apparently moved in straight lines. Scores on Stratton's battery were combined somehow to produce one global score. He was able to demonstrate that five of the six lowest scoring cadets subsequently failed flight training. While Stratton was conducting his tests, Francis Maxwell, also a member of the APA/NRC "Psychological Problems of Aviation including Examination of Aviation Recruits" committee, administered some or all of the MIT battery to another group of 44 applicants at Essington Field. In early 1918, a second group at MIT was tested.

Henmon (1919) indicates that a total of 40 different tests were administered in these four data collection efforts. No description of the other 17 tests could be located. Data from all of these tests with the corresponding flight training records were sent to Thorndike for analysis. Based on Thorndike's analysis, six tests were selected for further study. Three tests-- simple visual reaction time, simple auditory reaction time, and the emotional stability test—were from the French and Italian selection batteries. Two tests—the postural swaying and equilibrium tests—were from the original battery of 23 tests. A variation of the equilibrium test was added, which required the candidate to respond to sudden changes in tilt versus detecting gradual changes in tilt.

Developing Pilot Selection Tests—Spring, 1918 to the Armistice

The Army stopped all pilot applicant examinations on February 9, 1918 because the number of successful pilot applicants waiting to begin training greatly exceeded the capacity of the training facilities. Applicant testing was never resumed because the Armistice was signed before the backlog of pilot applicants had been sufficiently reduced. Several articles (e.g., Henmon, 1919; Stratton et al., 1920) imply that investigators assumed that pilot selection would resume as soon as the training facilities had cleared the backlog of cadets. Consequently, the Army continued research on pilot selection up to and shortly after the Armistice in November, 1918.

In the spring of 1918, Stratton began supervising two studies conducted at different flight schools (Henmon, 1919). In the first, he and Henmon administered a battery of nine tests to 150 cadets and flight instructors at each of two training schools, Kelley and Rockwell Fields. This battery consisted of the six tests identified by Thorndike: Stratton's curve test, the Mental Alertness Test, and the Thorndike-Kelly Athletic Interests Inventory.

At each school, 50 of the cadets were highly rated, 50 were poorly rated, and 50 were of unknown ability. Some of the poorly rated cadets were already grounded at the time of the testing because of unsatisfactory progress in the training program. All of the highly-rated and poorly-rated participants were tested twice on successive days to determine the reliabilities of the tests, which are not reported. Scores on the tests were correlated with ratings of flight ability obtained from the officers in charge of each stage of flight training. Apparently, instructors also were rated, and their data were combined with those of the cadets.

Stratton's curve test, the two simple reaction time tests, and the new version of the equilibrium test all had non significant correlations with ratings. The startle and the equilibrium (perception of slow changes in tilt) tests had the highest correlations with rating of ability (r = .26 for both, p<.05) followed by the Mental Alertness test (r =.23, p<.05). No data are reported for the Thorndike-Kelly Athletic Inventory because the test was already adopted for use. Henmon also reports a correlation of .7 between ratings and a composite scores based on the Mental Alertness Test, the Thorndike-Kelly Athletic Inventory, the emotional stability test, and the equilibrium test.

Henmon then identified for the flying command the bottom 10% of the cadets and the top 4%. His identification of both the worst and the best of the cadets proved accurate enough for the Army to approve the use of the six tests in the pilot selection battery. However, the Armistice was signed by the time the equipment had been purchased and the test administrators trained.

The second study supervised by Stratton (Stratton et al., 1920) was conducted at two additional flight training schools located at Taylor and Souther Fields and appears to have been conducted after the first study was completed. The conceptual approach for this study was substantially different from preceding research in that it began by identifying the basic abilities and personality traits needed for success in flight training. Then the method used to measure each ability or trait was identified. These methods included many of the non-medical selection instruments from the selection process as well as the six tests identified by Henmon (1919). Stratton et al. determined that several of the required abilities and traits were not assessed by any existing selection instrument. Consequently, five new tests were developed to measure these previously unassessed abilities and traits.

The dependent measure was based on instructor ratings of the cadet's dual and solo flying performance based on a 4-point scale. These scores were combined to give an overall score. At Taylor Field, the cadets were placed into one of 8 groups (performance categories) based on their overall score. A different rating system was used at Souther Field, which resulted in 14 performance categories. Fifty cadets were tested at Taylor Field and 70 at Souther Field. All tests were administered twice on successive days.

Three of the five tests may be related to spatial abilities. One test asked the cadet to determine where a parabolic curve, if continued, would intersect a horizontal plane. Scores on this test showed a low (r = .11 and .25 for Taylor and Souther Fields, respectively) but usable correlation with ratings from flight training. The second test required a judgment of relative speeds. This test showed a correlation of approximately .22 for both schools with ratings. Again, the magnitude of this correlation was considered to be usable. The third spatial test required the examinee to find his way through a finger maze. The cadet could see only a very small portion of the maze at any given time. Scores on this test did not correlate with ratings of flying ability. The fourth test measured grip strength. The results from this test showed no significant correlation with ratings.

The fifth test is most important because derivatives of it are still in use today. The test was constructed of used aircraft parts and consisted of a seat, rudder bar, and control stick. Stratton et al. called this test the "complex reaction time" test. Mashburn (1939) attributed this test to the Italians (see Appendix A) but gives no reason for this attribution. Henmon (1919) mentions that the APA/NRC committees had access to information about French and Italian selection instruments in 1917. Dockeray and Isaacs (1921) provide a detailed description of the Italian test, indicating that it was similar but lacked the rudder bar. It appears, then, that this test was derived from an earlier, successful Italian instrument.

At the start of each trial, the cadet sat in front of a screen on which was projected an arrow pointing either left or right and the letters B, F, L or R. The arrow indicated the direction that the rudder bar should be pushed; the letter indicated the direction the stick should be moved (backward, forward, left, or right). Reaction time was measured from the time either the rudder or the stick was moved until both movements were completed. Errors were also recorded. Each cadet performed 20 trials per day.

Because derivatives of this test are still in use today, the results should be examined carefully. The percentage of trials with an error ranged from 0% to 73%. Average reaction time ranged from 640 to 1640 ms. The correlation between average reaction time and performance category was 0.17, which was significant. However, Stratton et al. noticed a speed/accuracy tradeoff. The correlation between speed and accuracy was -.415 for cadets from Taylor Field and -.337 for those from Souther Field, both of which appear to be statistically significant. The authors wanted to correlate the cadet's average reaction time with his flight performance category, taking errors into account. The correlation between correct reaction time and performance category was .16 for cadets at Taylor Field, which may be non significant. The authors subsequently calculated a partial correlation of .26, which was statistically significant. Unfortunately, the authors provided no information on what was being partialled out of the correlation.

Stratton et al. (1920) reached four interesting conclusions from this study. First, practice had a significant effect on many of these tests. The authors recommend more practice before data collection. Second, studies of this nature need at least 200 cases to reach meaningful conclusions. Third, instructor ratings are contaminated by factors not related to flight performance. Fourth, ratings of flight performance suffer from range restriction.

Armistice through 1919

In September, 1918 a group of medical personnel and psychologists from the main research laboratory at Mineola, N.Y. were sent to the flying school at Issoudun, France to establish a research laboratory. After the Armistice was signed, these psychologists began a series of studies on exceptional instructors, chasse (fighter) pilots, and observers (Dockeray & Isaacs, 1921). For pilots, the psychologists tested simple reaction time and hand steadiness and related them to ratings of flying ability. Unfortunately, Dockeray and Isaacs do not indicate if the ratings were for preliminary, advanced, or mission-specific training.

All of the reaction time tests demonstrated low correlations (rs = .13 to .04). Hand steadiness was scored for the "appearance of tremor" and participants placed into one of five groups based on the amount of tremor demonstrated in the test. The correlation between scores on the hand steadiness test and ratings from flight training was .73.

Dockeray and Isaacs (1921) made several important observations based on their results and from observations at Issoudun. First, they felt that simple reaction time tests were inappropriate for selecting pilots. In contrast, they felt that complex reaction time tests that involved difficult discriminations had promise. They begun such experiments at Issoudun, but equipment and time limitations precluded the collection of useful data. Second, Dockeray began flight training to observe the personality traits of pilots. Based on his observations, he felt that individualism was the most important personality trait. Third, Dockeray made the prescient observation that intelligence was the most important factor for a successful pilot.

In 1919, the laboratory at Issoudun was closed and the staff moved to the main laboratory at Mineola, NY where they continued their research. Johnson (1920) provides a summary of work conducted during 1919 at this laboratory, the Air Service Medical Research Laboratory. Two items are noteworthy. First, preliminary studies identified two promising tests for pilot selection. One is a test of "the ability to control the co-ordinated activity of certain systems of muscles (p.

451)". This description could refer to the "complex reaction time test" described by Stratton et al. (1920). All of these investigators also had been moved to the laboratory at Mineola. Conceivably, it could also refer to the hand steadiness test described by Dockeray and Isaacs. The second test appears to be a discrimination reaction time test that was not described in any prior work. Conceivably, this test was one that Dockeray and Isaacs began at Issoudun. These tests were found to be correlated with estimates of flying ability obtained from instructor pilots.

The second item of note concerns grade inflation. Johnson remarked that flying grades are not sensitive to differences in ability. He found that 85% of the cadets at one flying training school were scored within 5 points of each other on a scale of 0 to 100 although he did not mention if this score was a daily grade or an overall score. He notes that this type of distribution makes it difficult to compare flying scores with other measures.

Track Selection

All flight cadets appear to have received the same ground school, primary, and advanced training curriculum although the type of aircraft used for advanced training varied from school to school. After graduation from advanced training, the pilots had to be selected for three specific missions: pursuit (chasse), observer, and bomber. Track selection appears to have been based on three pieces of information. The first, and perhaps the most important, was the final grade in flight training. According to McComas (1922, p.190), each instructor kept a daily record of his students' progress. This record also included notes about the student's intelligence, personality, and other traits the instructor noted. When the student completed training, he received a final grade. McComas does not indicate if this grade occurred after primary training or advanced training, but advanced training seems to be the logical choice. The "best" cadets (highest final grades) were recommended for pursuit. The next group was recommended for bombers, and the poorest were recommended for observers (McComas, p.191). The second piece of information used for track selection was the pilot score on the rebreather apparatus, which tested the pilot's resistance to hypoxia. A candidate for pursuit training had to score well on the rebreather test because much pursuit flying was done at high altitudes without any oxygen equipment. The third piece of information came from the psychiatric examination given as part of medical examination, which noted if a candidate was fidgety or phlegmatic. Neither type was considered good for pursuit pilots. Pilots who did not receive high enough final scores to be pursuit pilots but who were considered to be level-headed and reliable were considered to be bomber material.

Summary

Before the start of World War I, all pilot selection tests were medical examinations. Within 6 weeks after the declaration of war, the U.S. Army pilot medical examination had expanded to include biodata and a survey of athletic interests. A relation between athletic achievement and success in flying training was observed quickly. The athletic interest survey subsequently was expanded and refined to provide more information on a candidate's athletic interests and achievements. This relation between success in flight training and athletic interests and achievement has continued to the present.

The need for an in-depth assessment of the candidate's intelligence also was realized early, and an extensive intelligence test was added to the battery in 1918. The war ended before tests of simple reaction and vestibular function were implemented. Tests of spatial ability and multi-limb

co-ordination appeared promising and were under development when the war ended. Some preliminary attempts at track selection began but were not given serious consideration during this period.

Perhaps more importantly, by the end of World War I, investigators made three observations that are still valid almost 100 years later. First, no single test can be used to predict success in flight training (Henmon, 1919; Stratton et al, 1920; Thorndike, 1919); many factors need to be assessed. Second, flying grades are frequently poor criteria because of the lack of variability in the scores (Johnson, 1920), poor interrater reliability, and because other factors, such as bearing as a military officer (Stratton et al., 1920), affect the grades. Third, intelligence is important to success as a pilot.

BETWEEN THE WARS I

Background

After the Armistice, the uniformed psychologists who had been most prominently involved with research on pilot selection—V.A.C. Henmon, J.B. Watson, G.M. Stratton, F.C. Dockeray, and H.C. McComas—were discharged and returned to their academic positions and pre-war interests. These individuals were not replaced. The APA/NRC committees were dissolved quickly and, as a consequence, the U.S. military's efforts in pilot selection effectively vanished for a decade. Literally, no U.S. studies pertaining to pilot selection were published between 1920 and 1929 (except those cited in the previous section).

In the early 1930's limited research on pilot selection began again, in part because of the dismal state of military flight training. Razran and Brown's (1941) bibliography shows that the little U.S. pilot selection research undertaken during this period dealt predominately with medical issues, such as the effects of high altitude on cognition and the identification of neuroses; although a few studies dealt with personality and psychomotor issues. In the fall of 1939 the Civil Aeronautics Administration (CAA) began a large research effort dealing with the selection and training of civilian pilots. The military quickly became involved in this effort, and, beginning in 1941, the CAA became almost exclusively concerned with the selection and training of military pilots.

This chapter is broken into three parts. The first part reviews the limited information available on military pilot selection systems and their associated training results. The second part presents the few military studies published during this period that examined new selection tests. The third part deals only with the CAA research effort from the fall of 1939 to the end of 1940. Subsequent parts of this effort are described in the following chapter.

Military Pilot Selection and Training Results

Before discussing the selection systems, some mention should be made of the educational requirements for flight training. As noted in the preceding chapter, World War I applicants for Army flight training were required to have a high school degree. This requirement was dropped in 1920 and from 1920 to 1927, a pilot applicant could take an "equivalence" examination that was supposed to cover high school subjects (Guilford & Lacey, 1947, p. 46). In 1927, the educational requirement was raised to 2 years of college. However, again an equivalence examination was developed to allow those with less education to apply for flight training. The examination was retained until January, 1942 when the Army Air Force (AAF) Qualifying Examination became operational (Flanagan, 1947). No information on the Navy's educational requirements for this period was found.

The most striking aspect of the period from1920 to 1940 is the dismal performance of both the Army's and Navy's pilot selection systems. Five articles provide information on the failure rate in flight training with two of the five providing information on the actual selection process. In the first article Sutton (1930) examined the failure rate for Naval pilot training for a 12-month period from 1927 to 1928. For this year, the failure rate ranged from 87% for enlisted men who previously worked in non-aviation fields to 40% for officers with no prior flight training. Although this failure rate seems preposterously high, De Foney (1931) provides additional data

for Navy flight training from 1928 to 1930 showing a 55% failure rate. Sutton believed that many of the failures during this period were caused by poor assessment of personality but provided no data to substantiate his claim.

Davies (1940) describes traveling Navy pilot selection boards, that is, selection boards that traveled to smaller cities and towns looking for applicants(Carlson, 1939 mentions traveling Army selection boards). According to Davies, the selection process included three character reference letters (similar to those required in World War I), a physical examination, and a "neuropsychiatric or personality" portion that was a completely unstructured interview focusing on the candidate's childhood and school years. Between 41% and 82% of the applicants failed the cadet selection process. The successful candidates did not join the regular Navy and may have become reserve pilots.

In a review of the state of Army aviation, Mashburn (1939) mentions five important facts. First, physical standards had gradually increased since World War I. At the time this article was written, approximately 80% of the candidates failed the flight physical. The increased severity of the physical did not, however, result in a lower failure rate (p. 431). Second, as noted earlier, the educational requirements for flight training were raised from completion of high school to completion of 2 years of college. The increased educational standard did result in a lower failure rate (p. 433). The secondary effect, however, was a decrease in the number of applicants. Third, the failure rate in flight training was very high by modern standards. The March, 1927 class had an 87% failure rate. Over some unidentified period of time, the failure rate in flight training averaged between 60% and 65% with a minimum of 44% failing in a given class. These failures overwhelmingly occurred in flight training, not in ground school (Mashburn, 1935). Fourth, pass/fail from flight training had been accepted as the criterion to be used in selection. Fifth, the Army had not yet developed standardized rating scales for flight training. Several attempts had been made, but all were unsuccessful.

Finally, Flanagan(1942) presents failure rate data for the Army from 1924 to 1941. These data confirm Mashburn's (1939) information. They also show failures by training stage. Between 14 and 20% of the failures occurred in advanced training for the period from 1924 to 1927. In contrast, for the period from 1937 to 1940, only 1 to 2% of the failures occurred in advanced training. Nevertheless, the overall failure rate was still approximately 45% for this 3-year period.

The data from these five studies indicate that the military went backwards in terms of selection success from World War I. Mashburn remarks that the World War I tests were discounted because of the poor validation process, that is, some of the World War I tests were validated by comparing good cadets with poor cadets or good pilots with poor pilots. Mashburn's statements appear to apply to the apparatus tests, not to the written tests, such as the Mental Alertness Test because the written tests were validated by Thorndike using a predictive validity approach with performance in training as the criterion. This approach is still acceptable today. Unfortunately, Mashburn never describes the non-medical tests used between 1919 and 1939. Thus, it cannot be ascertained if any of the non-medical portion of Form 609 or apparatus tests were used during this period. Mashburn (1935), however, does comment that "no objective or scientific tests are used to measure potential aptitude in an applicant for flying training," which strongly suggests that even the written tests were dropped sometime after the end of World War I.

Military Research

The few psychological research studies conducted during this period were concerned with the development of apparatus tests. In 1925, Thorne developed the Thorne Reaction Time Apparatus (Mashburn, 1934a), which measured simple reaction time and discrimination reaction time. No studies were reported using this device. In 1927, the Complex Coordinator Test purportedly was developed by L. J. O'Rourke, Director of Personnel Research for the U.S. Civil Service Commission. This device consisted of a seat, "airplane-type controls" (most probably the control stick and the rudder bar) mounted in the same position as in an aircraft (Mashburn, 1934a), and an upright panel mounted in front of the candidate. Mashburn mentions a series of red, green, and white lights and a buzzer on the panel but provides no details about stimuli or responses except that the candidate made 62 responses that could involve one or both of the controls. How this test relates to the World War I Complex Reaction Time Task described by Stratton et al.(1920) is unclear; the descriptions of the physical devices are very similar. Mashburn cites the Italian version of the test (See Appendix A) but makes no mention of the American test. Given that O'Rourke had no apparent background in aviation, it is likely that this test was created without reference to prior work.

Mashburn (1934a) performed a preliminary validation study on the Complex Coordinator Test using data from 1,394 student pilots entering flight training between 1925 and 1931. For each student, the reaction times to the 62 stimuli were combined in an undisclosed manner to give an overall score. Mashburn notes that the score distribution was approximately normal. Mashburn then grouped the overall scores into 16 categories and determined the percent of students who passed advanced flight training versus the percent who failed flight training in each category. Approximately 74% of those in the highest category (shortest overall reaction time) graduated versus 14% of those in the lowest category. The Complex Coordinator Test had not been adopted as a selection instrument at the time this article was written (circa 1934).

Mashburn (1934b) developed a third apparatus in 1931 named the "Serial Action Apparatus". This apparatus had aircraft-type controls (control stick and rudder bar). A board in front of the candidate contained three sets of lights. Each set contained two parallel rows of 13 lights. One pair of rows was vertical, one was horizontal, and one was curved. For each pair, one row contained green lights; the other, white lights. The green lights were controlled by the experimenter; the white lights, by the candidate. Each row of white lights was controlled by a different movement of a control device For example, forward and back movements of the control stick changed which light was illuminated in the vertical row. "One or more" (p. 158) green lights could be illuminated at any given time. The candidate was to move the controls as quickly as possible to match the white lights to the illuminated green lights. As soon as the lights were matched, the next set of "one or more" lights was illuminated.

The Serial Action Test was developed because the Complex Coordinator Test required manual scoring of both correct responses and errors, which took too long. The Serial Action Test did not allow the candidate to progress to the next stimulus(i) until he had responded correctly to the current stimulus(i). This allowed the error reaction time to be incorporated into the total reaction time, simplifying scoring. Additionally, the operator of the Complex Coordinator Test had to

present each stimulus, which was time consuming. The stimulus presentation of the Serial Action Test was automatic; the operator only had to start the apparatus.

Glenn (1935) presents data from the Serial Action Test from 1466 individuals who had been selected for flight training. The data were analyzed similarly to those of Mashburn (1934a) but with 21 response time categories rather than 16. Only approximately 16% of those with total response times in the fastest category failed flight training. Of those in the three slowest time categories, approximately 79%, 90%, and 79% failed.

The other focus of research during this period was concerned with the "psychological" evaluation of naval candidates and is reported in two articles by De Foney (De Foney, 1931, 1933). The single study reported in the two articles began in 1928 with data presented for 1928 to 1930. The study involved the administration of a background questionnaire, a personality assessment, and tests of concentration, attention, and reaction time for 628 individuals who appear to have been selected previously for flight training. No details or names are provided for any of the tests. The primary purpose of all of the tests was to determine how an individual behaved when confronted with unfamiliar and difficult situations. The students were rated on a five-point scale by the medical officer of the flight training squadron on each of the following constructs: courage, stability, aggression, concentration, intelligence, and reaction time. Pass/fail from flight training was the criterion.

De Foney (1931, 1933) found that ratings of intelligence, reaction time, and concentration were not related to success in flight training. Of the 628 students tested, 416 were classified as poor candidates for flight training. Of these 135 passed flight training (32%). De Foney found that the accident rate for these 135 pilots was approximately twice that of Navy pilots as a whole in 1928-1929. De Foney does not indicate if these were training accidents or operational accidents. Consequently, one could argue that these data reflect only a higher accident rate for less experienced pilots.

The second study (De Foney, 1933), however, demonstrates that the selection system was identifying more accident-prone pilots. De Foney essentially repeated the methodology described in the 1931 article with an additional 677 students. In this study, DeFoney examined the accident rate for all of the aviators classified as good candidates in both studies versus those classified as poor candidates. The poor candidates had an accident rate approximately three times higher than that of the good candidates. At the conclusion of the second article, De Foney recommends more research and refinement of the personality assessment method.

CAA Research

In the fall of 1939, the Civil Aeronautics Authority (CAA) instituted a program to train 10,000 civilian pilots, which was quickly expanded to 50,000 (Viteles, 1945). The purpose of this program was to prepare young adults to fly private and commercial aircraft and, as a consequence, develop the U.S. light aircraft industry. The program was administered through universities that had civilian flying programs, and eventually 40 universities became involved in the research. Shortly after this program began, additional funding was given to the NRC for research on the selection and training of civilian pilots. The NRC established a committee, the Committee on Selection and Training of Aircraft Pilots (CSTAP), to oversee the research needed

to develop civilian flight training and identify young adults who could complete the training successfully. The committee included both psychologists and physicians from universities, the military, and branches of the federal government.

The CSTAP was in existence until at least 1945. The committee's work is documented in numerous CAA reports, none of which could be located. Consequently, information about the work of this group rests on two journal articles (Jenkins,1941; Viteles, 1945) and a book chapter (McFarland, 1953) although Poppen (1941) refers briefly to the results of the research. The CSTAP clearly was aware of the World War I work because one member of the committee, H. M. Johnson, had been the director of one of the laboratories concerned with pilot selection. In some situations, which will be described below, the CSTAP continued the World War I validation efforts on specific tests. Nevertheless, J. G. Jenkins, the chair of the committee, believed that much of the earlier work was limited because of the statistical techniques available at that time and the test development processes in use at that time, which were more typical of experimental psychologists than industrial psychologists. One of the biggest shortcomings of the World War I effort, in his opinion, was the lack of a job/task analysis for pilots. A job/task analysis for civilian flying was promptly conducted and used in several of the first studies. Unfortunately, Jenkins did not include details of the job/task analysis.

One of the earliest studies was concerned with identifying useful performance criteria. The investigators quickly found that civilian flight instructors demonstrated poor interrater reliability when assessing student performance, a fact that had been noted in World War I. The investigators developed detailed scales for scoring overall maneuvers as well as more global rating scales. A second early study also concerned student pilot evaluation. Prior to the CSTAP research, the maneuvers performed as part of check rides for pilot licenses were not standardized. That is, an applicant could be asked to perform one set of maneuvers for a private pilot license at one facility and a completely different set of maneuvers at another facility. One of the major contributions of the CSTAP was the development of a standardized set of maneuvers to be performed at any facility for a given license.

Sometime prior to December, 1940, the U.S. Army and Navy began submitting requests to the CSTAP for specific types of research. One of these topics concerned military flight training criteria. The CSTAP investigated washout rates in the Army and Navy and found them high enough to be useful. They also developed some type of assessment of pilot performance during combat although Jenkins (1941) is deliberately vague about the nature of the measure.

Between the fall of 1939 and late 1940, the CSTAP investigated several selection instruments. The first of these was a test of intelligence that was found to be predictive for both civilian and military flying. Jenkins remarks that no personality test was found to be predictive, but an interest inventory and a biographical inventory showed promise. His comments do not mention if these tests were for civilian or military selection. Based on comments from Fiske (1947) that will be discussed in the next chapter, it seems that these tests were studied in the context of naval flight training.

Based on the job/task analysis performed previously, several psychomotor tasks were investigated (Jenkins, 1941). Two showed good predictive results for both civilian and military

fight training, whereas one unnamed apparatus test from World War I had very low predictive validity for both types of flight training. Apparently the emotional stability test (Jenkins deliberately omits specific names and test descriptions) from World War I was examined as a predictor but no results are described. Again, Jenkins never describes the test population. Other medical tests, such as the tilt tests described in the preceding chapter, were refined.

In 1940, the CSTAP began to collaborate with the Navy on the development of selection tests for naval aviators to reduce the high washout rates noted earlier (McFarland, 1953). Because no job analysis was available for naval aviators, the initial selection battery consisted of 40 "psychological and physiological tests" (McFarland, 1953, p.40) chosen on the basis of the results obtained to date by the CSTAP and on expert opinion. The battery was given to 919 naval flight cadets. No cadets were eliminated based on the results of the tests battery; all were tracked through flight training. From the training performance data, a test battery for naval aviators was assembled consisting of the Wonderlic Personnel Test of Mental Ability (See Carlson, 1941 for results and suggested cutoff scores), the Bennett Mechanical Aptitude Test, and a biographical inventory. Several apparatus tests were promising, but administrative difficulties precluded their inclusion in the battery.

Summary

Useful selection instruments and promising lines of research identified in World War I were abandoned by both the Army and Navy shortly after the Armistice. Consequently, the period between the wars was marked by extremely poor pilot selection systems for both the services as evidenced by extraordinarily high failure rates in training. The small amount of military research conducted between the wars appears focused on psychiatric/personality issues rather than on cognitive and psychomotor issues.

In 1939, the CSTAP, part of the CAA, was created and tasked with developing selection and training methods for private and commercial pilots. Between the fall of 1939 and the end of 1940, the CSTAP performed a job/task analysis for private flying, developed rating scales for check rides, developed standardized maneuvers for use in check rides, and identified several tests with predictive validity for student pilots. Many of the early changes to the military selection systems were the result of research by the CSTAP and not outgrowths of recommendations by prominent military investigators such as Bigelow(1940) and Mashburn (1939).The activities of the CSTAP continued throughout World War II and will be described in the next chapter.

WORLD WAR II

Background

As previously discussed, the pilot selection system developed by the Army in World War I was essentially dismantled after the war. Over time, both the Army's and Navy's pilot selection systems were reduced to a few tests, much to the detriment of the services. Little research was conducted between the wars, and the few studies that were performed were concerned with personality/psychopathology and psychomotor performance.

When the United States declared war on Japan in 1941, both services had to expand their selection systems to accommodate the huge numbers of pilots needed by the war effort. They also had to improve the selection system to reduce the high failure rates in flight training, which were unsustainable in wartime. Although some collaboration between the services is evident, the Army Air Force and the Navy developed their selection systems for the most part independently. The development of the Army Air Force's selection system and its component selection instruments are well documented in the 19-volume series entitled *Army Air Forces Aviation Psychology Program Research Reports* and will not be discussed in detail here. Instead, this chapter will review the initial battery construction and describe changes to the battery and the selection process over time. Issues surrounding specific selection instruments and abilities also will be discussed. Very little documentation was located on the Navy's efforts. The few available reports are discussed at the end of this chapter.

Like the World War I documents some of the World War II reports and articles suffer from a confusion of terminology, with the same test having multiple names. However, these problems are less severe and less frequent than in the World War I documentation. Unlike the World War I period the psychological terminology is essentially modern, physiology and psychology were distinct disciplines, and the statistical techniques and terminology used for data analysis are still in use today.

Because of security issues, very few articles on pilot selection were published during the war. Thus, this section includes work dealing with the war effort that was published through 1949. Research that was conducted between 1945 and 1950 is included in the following chapter.

Army

Staffing and equipment. As in World War I, many prominent psychologists became involved with pilot selection. In June, 1941 the Army established a Psychological Research Agency in its Medical Division. The first head of the Agency was J. C. Flanagan. In consultation with the National Research Council, he appointed directors of three Psychological Research Units and approved direct commissioning of many professors of psychology and other prominent psychologists for positions in these Research Units. The Units initially were staffed with 44 officers and 200 enlisted men, many of whom had master's or doctoral degrees in psychology (Flanagan, 1948). Among the prominent psychologists who initially accepted senior positions at the Research Units were J. P. Guilford and A. Melton. Many of the junior officers and enlisted men--such as R. L. Thorndike, Neal Miller, S. Bijou, L. G. Humphries, W. F Grether, J. E. French, R. M. Gagne, and P. Fitts—began their illustrious careers working in the Research Units. Thus, again the United States was extremely fortunate to have many of the best psychologists of the period involved with aircrew selection.

In World War I the first experimental battery of 23 apparatus tests was administered in June, 1917. Why these 23 tests were selected for the initial battery is unclear; no rationale or scientific source for these tests could be located. The physical source of the tests also was never identified. In World War II, the source of the initial apparatus tests is clear; they were borrowed from the laboratory of R. H. Seashore at Northwestern University. As Melton (1947, p. 9) notes about the October 1941 battery, "It may be fairly stated that the tests were employed without reference to any particular hypothesis regarding the psychomotor tests most likely to predict success in pilot training." Between January and June 1942, apparatus tests continued to be borrowed from Columbia University, Yale, and the University of Missouri as well as from Northwestern University. By July 1942, the Psychological Research Units and their associated facilities could produce sufficient apparatus tests for mass testing.

Criterion issue. One of the most serious issues confronting the developers of any selection system is the criterion. Some of the documents describing the Army Air Forces (AAF) selection and classification system argue that the most appropriate criterion for military pilots is success in combat flying (e.g., Kaufman, 1943). However, a document describing the development of the AAF Qualifying Examination (AAFQE) gives three reasons why this criterion was not adopted for either the AAF pilot selection program or the classification program (Office of the Air Surgeon, 1944). First, rating men's combat performance is extraordinarily difficult. Data simply are difficult to obtain under combat conditions, and outcomes are affected by many factors other than the skill of the pilot. Second, months separated test administration from the collection of combat data. Using combat data as the criterion would have slowed down the development of selection instruments. Third, the selection system initially consisted of only one instrument, the AAFQE, which was designed to replace the educational requirement. Prediction of combat performance was not, therefore, an appropriate criterion for this instrument. The Aircrew Classification Battery, which consisted of printed and apparatus tests, was designed to be used with a candidate's preferences to assign the candidate to one of three specialized roles: pilot, navigator, or bombardier. Arguably, therefore, combat performance was not an appropriate criterion.

Training performance became the criterion by default. The use of performance in training had recognized drawbacks, but, as noted by Melton (1947), training measures were the most accessible and could be obtained in 5 to 6 months after testing. Additionally, training was conducted under more standardized conditions than to combat.

The test weights were developed primarily to predict pass/fail from elementary flight training, the first of the three flight stages (elementary, basic, and advanced) (Melton, 1947, p. 55). Performance in elementary training was given the most weight in assessing the predictive validity of a test because most failures occurred in elementary training and because the flying staff made very carefully considered decisions to fail a student from this stage. Most printed and apparatus tests were evaluated exclusively against pass/fail from elementary training.

Little information was located on failure rates from elementary training. Melton (1947, p. 57) reports failure rates by geographical area for early 1943. These rates ranged from 14% to 43%. Deemer and Rafferty (1948) remark that failure rates could range up to 50% per class. Ground

school performance, which occurred before elementary training, was never used as a criterion for evaluating a test probably because of the low failure rate (Kaufman, 1943). The Psychological Research Units did attempt to develop other criteria for evaluating the selection and classification tests. This included the development of various rating scales and objective measures of flight performance. None of these efforts were successful enough to be used operationally.

Selection process. The selection process varied over the course of the war. Major changes are described in each section below.

***December 1940-December 1941*.** As noted in the previous chapter, no documents describe in detail the Army selection system used between 1920 and 1940. Nevertheless, several important facts are known about the selection system in use at the beginning of World War II. First, between 1927 and January, 1942, all applicants for flight training must have completed at least 2 years of college or passed an educational equivalence examination covering nine college subjects (Flanagan, 1947). The equivalence examination used an essay format and was difficult to score. Second, applicants were accepted into flight training on the basis of an "adaptability for military aeronautics" rating (Flanagan, 1948). These ratings were based on interviews conducted by flight surgeons. In these interviews, the flight surgeons obtained information on the candidate's personal and medical history, as well as interests and rated the candidate on 21 personality attributes (see Deemer & Rafferty, 1948 for examples of rating forms). Flanagan indicates that flushing and hand tremor were also rated, but how flushing and tremor were induced is not stated. Flanagan (p. 85) mentions an analysis conducted on the interview forms that identified five major contributors to the rating: education, vocational achievement, interest in flying, national origin, and family income. No other information on these interviews was found.

By November 1941 the US Army realized that they could not meet the manpower demands of the war with the existing educational requirements. They also realized that they could not process the required number of applicants if each applicant had to be interviewed by a flight surgeon; there were simply too few flight surgeons to do the interviews. Consequently, the Office of the Air Surgeon began work in December 1941 on an intelligence test to be used in lieu of the educational or educational equivalence requirement (Flanagan, 1948, p. 21). This test became the AAFQE.

***January 1942-July 1942*.** Sometime between the end of 1941 and early 1942, the Army began to establish several hundred Aviation Cadet Examining Boards. The purpose of these boards was to administer the selection system for aircrew. These boards were set up around the country, but no information about specific locations was found. In World War I, these were staffed by an officer who was responsible for the day-to-day activities of the center and volunteer physicians from the local community. No comparable information was found on the World War II Examining Boards except that few, if any, professional psychologists were associated with these boards.

A candidate for an aircrew position presented himself at one of these boards with three letters of recommendation (Flanagan, 1942, p. 232). He completed an application and was given the AAFQE, which became operational in January, 1942, and a general military physical

examination. The AAFQE was developed for easy administration and quick, on-site scoring to eliminate the need for professional psychologists (See Deemer, 1947 for detailed information on test administration). Thus, it appears applicants knew immediately if they failed either examination.

Next, the board determined if the applicant had a satisfactory moral character. No information was found on how this determination was made or how long it required. If the candidate passed both examinations and had a satisfactory moral character (Flanagan, 1942), he became an aviation cadet.

Successful candidates then proceeded to a regional examining center where they were given a flight physical examination. Those who failed were assigned to ground crew specialties. Those who passed the examination were assigned to training either as a pilot, a navigator, or a bombardier. Flanagan (1948) describes a short (three-test) battery used to select navigators and bombardiers but not pilots. It is not clear how the pilots were separated from the navigators and bombardiers.

***July 1942-August 1945*.** In July 1942 the Aircrew Classification Battery became operational and was administered at the regional center after the second physical examination. The purpose of this battery was to assign a candidate to one of the three aircrew specialties. Thus, the Aircrew Classification Battery eliminated the three-test examination for selecting navigators and bombardiers.

The test scores from the battery had to be combined in such a way that the candidate's aptitude for each of the three aircrew specialties could be compared directly. This was done by assigning each test in the battery three different weights: one for pilot aptitude, one for navigator, and one for bombardier. Correspondingly, three composite scores were created for each candidate: one representing the candidate's aptitude as a pilot, one as a navigator, and one as a bombardier. Each composite was created by multiplying the candidate's score on a given test by the weight for one of the three positions. The weighted scores for each test then were summed to create, for example, a pilot composite. This composite then was transformed to a stanine score and the process repeated with the navigator and bombardier weights. The candidate was assigned to a specialty based on the candidate's stanine score for each aircrew category and his preference. It is important to note that no one failed the Aircrew Classification Battery.

Beginning in August 1942, the Aircrew Classification Battery began to be used for selection as well as for classification. A minimum stanine scores was established for navigators. In December 1942 minimum stanine scores also were established for pilots and bombardiers. These cutoffs were changed throughout the war in response to manpower needs and training data. Processes were also established for dealing with candidates who scored below the cutoffs on all of the aircrew specialties. The number of aircrew specialties gradually was expanded to seven, including fighter pilot and bomber pilot.

Development process for selection and classification tests. During the summer of 1941, the staff of the Psychological Research Units began reviewing the pilot selection reports from World War I and the 1930's (Flanagan, 1948), including those conducted for the UK Royal Air Force

(RAF) and the Royal Canadian Air Force (RCAF). Reports concerning apparatus tests were given special attention. During the summer of 1941, personnel from the Psychological Research Units reviewed classification tests used by the UK RAF and visited an RCAF selection facility in Toronto to observe operations and discuss selection methods. At this time the RCAF was using the Link Trainer as a classification instrument. This visit led the Americans to conclude that the Link Trainer was not a feasible selection instrument for the AAF because of the time required to administer the test and because simpler apparatus tests gave comparable results (Melton, 1947, p. 11).

Sometime in 1941 the staff began receiving reports of Flying Board proceedings on student eliminations. These reports usually included statements from both the instructor and the student about why the student had not made sufficient progress in flight training (Melton, 1947, p. 61). Based on initial results from the Flying Boards, the staff began to categorize the causes of failures and preliminarily identified 20 categories of failures, which are shown in Table 1. During the summer of 1941, a decision was made to develop tests to assess skills, abilities, traits, and interests that were reflected in the 20 categories (Flanagan, 1948, p. 13). In 1942, after the staff obtained reports of the Flying Board proceedings for 1,000 student eliminations from flight training, the groups shown as headings in Table 2 were developed. Each category subsequently was assigned to one of four groups. The apparatus tests specifically were designed to assess those causes that later were listed in the "coordination and technique" group. The decision was also made to try out the tests as soon as they were developed, resulting in a continuous process of test development, refinement, and validation that continued throughout the war.

Table 1. Reasons for elimination from flight training

Intelligence and Judgment	**Coordination and Technique**	**Alertness and Observation**	**Personality and Temperament**
Judgment	Coordination	Visualization of flight course	Absence of tenseness
Foresight and Planning	Appropriateness of controls used	Estimation of speed and distance	Confusion and nervousness
Memory	Feel of the controls	Sense of sustentation	Fear and apprehension
Comprehension	Smoothness of control movements	Division of attention	Temperament
	Progress in developing technique	Orientation	Motivation and attitude
		Speed of decision and reaction	

From Melton (1947, p. 62)

No formal job analysis (as the term is understood today) was conducted to identify the knowledge, skills, and abilities needed for success as a pilot. Nevertheless, the Psychological Research Units did conduct additional studies throughout the war to identify the characteristics of successful pilots. Three of these bear mentioning. One study required flight instructors of failing students to rate the student on the 20 categories listed in Table 1. In the second study, the flight instructors made comments on a grade slip about weaknesses or problems the student encountered. These comments were categorized into a "more suitable form for a job analysis

study" (Guilford & Lacey, 1947, p. 3), presumably into the 20 categories listed above. In the third study, officers in charge of combat squadrons rated the 20 categories on a 9-point scale for the minimum acceptable level for combat operations. The results of these studies guided test development during the war.

Before the August 1942 battery became operational, individual tests were administered to candidates on an experimental basis. An important question concerns how the original battery was assembled. That is, what guided investigators in their choice of tests for an initial experimental battery?

Melton addresses this issue for the apparatus tests in two articles (Melton, 1943, 1947). Certain details in these articles contradict each other. In the 1947 article, Melton states that from January to July 1942 the investigators at the Psychological Research Units were evaluating a group of 12 apparatus tests that had been borrowed from psychology laboratories at Columbia, the University of Missouri, Northwestern, and Yale (p. 12). The tests were simply assembled ad hoc. During the evaluation period, the investigators determined the predictive validity of the tests and studied practical aspects of testing. Those tests with good characteristics became part of the first operational battery shown in Table 2. Two tests in the initial experimental battery, Steadiness and Finger Dexterity, were not in the group obtained from university laboratories. The Finger Dexterity Test was included in the experimental battery because investigators assumed *a priori* that it was a necessary ability for bombardiers and because the equipment was simple to build (p.4). The Steadiness Test may have been included for similar reasons. No tests in the experimental battery were given a zero weight in calculating the pilot stanine. Consequently, these tests were included in the first operational battery. However, both tests were eliminated quickly from calculating the pilot stanine as shown in Table 2.

AAFQE. A few comments concerning the AAFQE are in order. The AAFQE was a selection instrument that was designed "to permit the preliminary screening out of applicants for aviation cadet training who would have only a slight chance to succeed in training" (Davis, 1947, p. 3). Thus, a large percentage of applicants passed this examination. Because most failures occurred in the first stage of flight training (elementary), the criterion for item development was pass/fail from elementary flight training.

Applicants who failed the AAFQE could retake the examination an unlimited number of times. The only requirement for retesting was a 30-day waiting period. Because of this retesting policy, the Psychological Research Unit produced 17 versions of this test (Flanagan, 1948) between 1941 and 1945. The need for new test versions allowed promising research results to be incorporated quickly into the operational test.

The first version of the AAFQE became operational in January, 1942. The initial examination assessed six areas. These areas with their associated number of items were reading comprehension (15 items), contemporary affairs (30 items), mechanical comprehension (15 items), general vocabulary (45 items), practical judgment (15 items), and mathematics (30 items). The Psychological Research Unit constantly tried to increase the predictive validity of the examination. Consequently, the topical areas assessed by the AAFQE and the number of items per area changed extensively over time. The final version of the examination contained 15

reading comprehension items, 50 items assessing interest in aviation and general knowledge, 60 mechanical comprehension items, and 25 hidden figures items (Flanagan, 1948, p. 55).

Different methods were used to calculate the predictive validity of the 17 versions of the examination, making comparison difficult. Nevertheless, some feeling for the change in predictive validity can be obtained by comparing three versions of the examination. The first version of AAFQE showed a predictive validity of r_{bis} =.20 for the total score to pass/fail from advanced flight training. A year later, the content of the AAFQE had changed substantially, reflecting an increased emphasis on selecting pilots rather than navigators or bombardiers (Davis, 1947, p. 38). This version of the examination was administered to an unrestricted sample and showed r_{bis} =.46. The final version of the AAFQE had an r_{bis}=.39. By the end of the war, the consensus was that no more substantive increases in the predictive validity of the AAFQE would occur unless apparatus tests were administered with the examination (Davis, 1947, p. 51).

The AAFQE was considered to be a power test. Applicants were permitted a maximum of 3 hours to complete the test. Although the overall test had a time limit, it had no internal time limits. That is, the candidates themselves determined how much time to spend on each area.

Aircrew Classification Battery. From the fall of 1941 to August 1945, the Psychology Research Units were tasked with developing classification instruments. Briefly, this responsibility entailed the constant development and validation of new instruments with continuous refinement of existing instruments to increase their predictive validity and ease of administration. The constant experimentation and refinement led to the construction of new batteries, which became operational periodically over approximately a 3-year period.

Three primary sources of information—Flanagan (1948), Melton (1947), and Guilford and Lacey (1947)—describe the development of the classification battery and its component tests. Surprisingly, the information provided in these three documents is sometimes contradictory and often incomplete. The information describing the development of the first classification battery is particularly problematic. Flanagan (1948, p. 64), in the overview volume of the *Army Air Forces Aviation Psychology Program Research Reports*, indicates that the first classification battery became operational in July, 1942. Melton (1947), who authored the volume on apparatus tests, gives an August, 1942 date. Guilford and Lacey (1947), the authors of the volume on printed tests, provide no information on the development or content of the initial battery (in contrast to specific tests). Both Melton and Flanagan agree that six batteries were fielded—July/August 1942, December 1942, July 1943, November 1943, September 1944, and June 1945. Both Melton and Guilford and Lacey list the specific printed and apparatus tests that comprise each battery except for the June 1945 battery. Only DuBois (1947) describes the tests comprising this battery.

The names of the tests included in each version of the battery are shown in Table 2. Only the tests that were used for pilot classification are included in Table 2. For example, the Finger Dexterity Test initially was used in calculating the pilot composite. However, after approximately a year, investigators found that scores on this test had no predictive validity for pass/fail from flight training. Consequently, the test was retained in the battery, but its scores were only included in the navigator and bombardier composites. As discussed earlier, tests were

frequently modified to improve the ease of administration or improve reliability. Thus, only the generic test name is given in Table 2. The reader should consult either Melton (1947) or Guilford and Lacey (1947) for the test version used in a specific battery.

The printed tests required between 6 and 8 hours of testing time. The time required to administer the apparatus tests was strictly limited to 90 min., with 15 min. allocated for each test (Staff Psychological Reseach Unit #1, 1945). One of the immediate practical problems with administering the apparatus tests concerned order effects. Administering the apparatus tests in exactly the same order to all of the candidates was a logistical problem because large numbers of candidates had to be tested in a relatively short period of time. To maximize the use of the apparatus, the Psychological Research Units set a sequence in which the tests were to be administered. However, a given candidate could start at any point in the sequence (Melton, 1947, p. 37). Candidate's scores on specific tests were analyzed as a function of the ordinal position of the test (first, second, etc.). The results demonstrated that only some tests were affected by ordinal position and that this effect usually was limited to the first position. That is, candidates performed more poorly on a given test only when it was performed first. Other data suggested that individual tests were affected by the immediately preceding test. That is, transfer of training occurred between certain tests but not between others.

Table 2. Printed and apparatus tests comprising the aircrew selection battery

Test		**Battery**				
	August 1942	**Dec. 1942**	**July 1943**	**Nov. 1943**	**Sept. 1944**	**June 1945**
Printed Tests						
Technical Vocabulary	x	x				
Reading Comprehension	x	x	x	x	x	x
Mechanical Information		x			x	
Mechanical Principles		x	x	x	x	x
Mechanical Comprehension	x					
Mathematics A	x	x	x	x		
Mathematics B		x	x	x		
Numerical Approximation	x					
Numerical Operations A	x	x	x		x	
Numerical Operations B	x	x	x		x	
Arithmetic Reasoning	x					
Dial and Table Reading	x	x	x	x	x	x
Speed of Identification	x	x	x		x	x
Spatial Orientation I	x	x	x	x	x	x
Spatial Orientation II	x	x	x	x	x	x
Biographical Inventory			x	x	x	x
Practical Judgment					x	x
Arithmetic Reasoning					x	
General Information[1]		x	x	x	x	x
Instrument Comprehension I				x	x	x
Instrument Comprehension II				x		
Apparatus Tests						
Discrimination Reaction Time	x	x	x	x	x	x
Steadiness	x					

Complex Coordination	x	x	x	x	x	x
Two-Hand Coordination	x	x	x	x	x	
Two-Hand Pursuit						x
Rotary Pursuit[2]		x	x	x	x	x
Finger Dexterity	x	x				
Aiming Stress		x	x			
Rudder Control				x	x	x

Notes. From Guilford and Lacey (1948, p. 801-803). [1] This test was called "Technical Vocabulary" in the December, 1942 battery. [2] A divided attention attachment was added to the Rotary Pursuit Test beginning with the July, 1943 battery.

A few comments about the Complex Coordination Test are needed. The Complex Coordination Test shown in Table 2 is not the Complex Coordination Test developed by O'Rourke in 1927. It is the Serial Action Apparatus, an improved version of the Complex Coordination Test, developed by Mashburn in 1931. Perhaps because it underwent continuous refinements throughout the war, the Complex Coordination Test predicted as well, if not better, than it did in the early 1930's (Melton, 1943) despite major changes in the educational requirements and in the selection process. See the preceding chapter and Glenn (1935) for more information.

Unfortunately, the predictive validity data for both the printed and apparatus tests are presented for each test separately by battery version by testing location (Guilford & Lacey, 1947; Melton, 1947). That is, the predictive validity for the Complex Coordination Test, for example, is presented for the July 1943 test battery given at the test center in San Diego. Occasionally, data are cumulated over testing sites, but rarely over testing periods. Data for special groups such as West Point Cadets are presented in additional tables. Thus, it is extremely difficult to determine a representative predictive validity for a given classification instrument.

A summary of the activities of the Psychological Research Units (Staff of the Psychological Section, 1945) mentions that the battery developers were just beginning to develop two sets of weights for pilots: one for bomber pilots and one for fighter pilots. At the time the summary was published, work on the development of the two sets of weights was just beginning. Guilford and Lacey (1947, pp. 9-11) present a table with the average rating of 20 categories (See Table 2 for a list of the categories) by supervisors of combat teams. These ratings are presented separately for fighter and bomber pilots. Some of these attributes were assessed by tests included in the aircrew classification battery, such as mechanical comprehension and reading comprehension. Guilford and Lacey provide no indication that these ratings were considered for pilot track selection.

Unselected airmen study. One recurring problem with pilot selection concerns range restriction. Investigators involved with pilot selection frequently must determine the predictive validity of a new selection test. Usually, the sample available for making this determination has already completed part of the selection process. Because some of the candidates have been eliminated, the remaining sample is said to be "range restricted." The predictive validity obtained from the restricted sample typically underestimates the predictive validity of the test in an unrestricted sample. Estimates of the predictive validity for an unrestricted sample currently can be estimated through statistical methods. These statistical methods require knowledge of the score distribution of the candidates without any prior selection process, information that is frequently

not known. In such cases, the statistical correction cannot be conducted and the investigator must use the predictive validity from the restricted sample.

In World War I the effect of range restriction on validity scores was not appreciated. By World War II, the effect was understood, A summary of the activities of the Psychological Research Units (Staff of the Psychological Section, 1945) notes that the predictive validities of the classification tests were underestimated because the candidates were previously selected on the basis of the AAFQE and on the basis of their minimum composite aptitude score. The summary further indicates that the reported predictive validities were uncorrected because no adequate statistical techniques were available.

Because of the cost and administrative issues surrounding the AAFQE and the Aircrew Classification Battery, the Psychological Research Units needed to demonstrate the usefulness of the tests. To do this, they performed an "unselected airmen" study that would produce predictive validities for the AAFQE and the Aircrew Classification Battery on an unrestricted sample. In this study, which began in June 1943, 1,143 men were accepted into flight training regardless of their score on the AAFQE and the classification battery. All, however, passed the physical examinations. Of these 1,143 men, 878 failed to graduate from advanced flight training and become rated, a failure rate of 77%.

Not all of these were eliminated for flying deficiencies; 187 were physically disqualified or disqualified for administrative (fear of flying or personal request) reasons. Consequently, a more accurate representation of this study is that 700 out of 956 cadets failed flight training. The failure rate, 73.2%, is comparable to that observed between the wars and discussed in the preceding chapter.

Davis (1947) reports a predictive validity of r_{bis} = . 46 to pass/fail from preflight to elementary flight training for the total score on the AAFQE. Of the original group (1143 candidates), 42% had failing grades on the AAFQE. Flanagan (1948) only gives the predictive validities for the six printed tests that were most heavily weighted in the pilot stanine. These validities and the predictive validities of the apparatus tests are shown in Table 3. The predictive validities are for pass/fail from preflight through advanced training. The Aircrew Classification Battery (all of the printed and apparatus tests used to calculate the pilot stanine) had a predictive validity of r_{bis} = .66 with pass/fail from flight training.

Table 3. Predictive validities for some of the Aircrew Classification Battery for unselected airmen study

Test	Predictive Validity
Printed Tests	
Mechanical Principles	.43
Spatial Orientation I	.40
Spatial Orientation II	.34
Biographical Inventory	.33
General Information	.51
Instrument Comprehension II	.48
Apparatus Tests	
Discrimination Reaction Time	.42
Complex Coordination	.41
Two-hand Coordination	.36
Rotary Pursuit	.31
Finger Dexterity	.18
Rudder Control	.40

Note. From Flanagan (1948)

Specific abilities and traits. Because of the volume of documentation pertaining to the AAF program, only a few additional comments are necessary. Personality and motivation have been two areas of concern for selection since World War I. The reports of student eliminations described earlier were used to identify personality traits that appeared important to success in flight training. To assess the usefulness of these traits, the Psychological Research Unit tested 11 personality inventories, three of which were developed by the Unit, and four preference inventories (Guilford & Lacey, 1947). The Strong Vocational Interest Blank for Men and the Minnesota Multiphasic Personality Inventory were among those instruments tested. Neither one of these instruments was found useful for predicting pass/fail from flight training. The Unit also conducted interviews and performed direct observation of candidates performing various tasks. Again, neither of these methods produced acceptable predictive validities.

Motivation was assessed through general information tests and through a sports-and-hobbies-participation test. Both types of instruments correlated significantly with pass/fail from flight training. Because both the CSTAP and the Navy found positive correlations between scores on biographical inventories and success in flight training, the Unit developed a biographical inventory for use in January 1942. The results were encouraging and resulted in a series of refinements to the inventory and its eventual operational use in July 1943.

One test in the Aircrew Classification Battery was designed to assess timesharing--the Rotary Pursuit Task with Divided Attention. The Rotary Pursuit Task was added to the Classification Battery in December 1942. The divided attention attachment was first added to the test in July, 1943. The divided attention attachment consisted of a 2-alternative choice reaction time task. The apparatus consisted of two lights (the stimuli) and a push button located below each light. This apparatus was placed to the left (on the right for left-handed candidates) of the Rotary Pursuit Task. The dependent variable for the Rotary Pursuit Task was time on target, which was accumulated only when the candidate pressed and held the correct button (Melton, 1947).

The candidate received 5, 20-s test periods on the Rotary Pursuit Task alone, followed by an additional 10 trials under the divided attention conditions. A test-retest study was conducted on The Rotary Pursuit with Divided Attention Test. The average retesting interval was 28 days for 690 candidates. The uncorrected test-retest reliability was r_{bis} =.74. The predictive validity of the Rotary Pursuit Task with Divided Attention for pass/fail from elementary training for a group of 1,212 candidates was r_{bis} =.19, which is not significantly different from predictive validity for the Rotary Pursuit Task alone r_{bis} = .25 (N = 624).

The unselected airmen study described earlier also included the Rotary Pursuit Task with Divided Attention. The predictive validity of the Divided Attention Task for pass/fail from elementary flight training was r_{bis} = .29. Interestingly, the test retained some predictive validity to the more advanced stages of flight (r_{bis} = .21 (N= 363) and r_{bis} = .11 (N = 280)) for pass/fail from basic and advanced training, respectively). The test was also found to have a relatively low correlation with AAFQE, *r* = .20 and only .05 for pass/fail from ground school (sample sizes are not given but appear to be approximately 1,000). The general impression from Melton (1947) is that the Rotary Pursuit Test with Divided Attention was considered to be sufficiently predictive to be retained in the classification battery but was not as predictive as other tests, such as the Complex Coordination Test.

Navy

Staffing. The Navy's approach to the development of pilot selection tests was considerably different from that of the AAF. Unlike the AAF, there was no one, large organization staffed by tens of Ph.D.-level psychologists and supported by several hundred masters- and bachelor-level psychologists who were tasked with developing selection tests for pilots. Instead, the Navy had approximately 100 commissioned psychologists who were employed as counselors for aviation cadets, test administrators, and resource personnel for flight instructors (see Jenkins 1945, 1946). They trained interviewers for the selection boards and developed data recording forms. They also worked on selection systems for instructors and aerial gunners as well as for pilots (Fiske, 1946). Because these psychologists were assigned to many different types of activities, relatively few individuals were concerned exclusively with test development for the pilot selection battery.

As noted in the prior chapter, the Navy became involved with the CSTAP sometime in 1939, and much of the initial pilot selection test development was performed by the CSTAP at Navy flight schools. These efforts were described primarily in CAA reports that could not be obtained although Jenkins (1941) and Viteles (1945) provide comments about the CSTAP's support of specific projects for the Navy. Unlike the extensive documentation undertaken by the Army Air Force, no Navy technical reports were located that described either test development or changes to the selection system. Consequently, the changes to the selection process throughout World War II were pieced together from the few journal articles that could be obtained.

Criterion. Naval flight training appears to have been structured differently from that of the AAF (Jenkins, 1946; Norman, 1947). Training began with the Naval Flight Preparatory School consisting of ground school and physical conditioning. This stage was followed by the War Training Service stage during which the cadet continued to receive ground school training while

learning to fly a simple aircraft. The third stage was the Naval Pre-Flight School, which again consisted of physical training and ground school. The fourth stage was Naval Air Station Training. During this phase, the cadet learned to fly a primary trainer aircraft. The final two stages were advanced and carrier training.

Selection process. Jenkins (1946) provides the only description of changes to the pre-war pilot selection process. Unlike Davies (1940), Jenkins indicates that in 1940 naval pilot selection was conducted by selection boards at naval training bases with the local flight surgeons performing the physicals. The flight surgeons also supervised the administration of three paper-and-pencil selection tests, which began to be administered in early 1941 for validation purposes only. By the summer of 1941, however, the number of applicants had increased to such an extent that flight surgeons could no longer perform the physicals and administer the tests. Consequently, psychologists were recruited and commissioned to help with the test administration.

Jenkins's (1946) dating is unclear, but apparently beginning in mid-1941, the selection process occurred in four steps. In the first step, background information and basic physical data (height, weight, etc.) were obtained. Next, the flight surgeon administered the flight physical. If the applicant passed the physical, he was given the three psychological tests. The psychologist hand scored the tests and, if the applicant passed all three tests, he proceeded to the interview process. The description of the interview process is unclear; either one many-on-one interview was conducted or several one-on-one interviews were performed. Jenkins does state clearly that line officers, many of whom were World War I naval aviators, unstructured conducted the interview(s).

The psychologists assigned to the boards quickly realized that the sequence of selection instruments was inefficient and began changing the order of the four stages. In a few cases they introduced additional interviews into the process. These changes were instituted by individual psychologists and resulted in local variations in the selection process. At some unspecified time, the order of testing was standardized across all selection boards (Jenkins, 1946, p. 46). Additionally, individual psychologists worked with the interviewers at their local boards to develop a standardized interview. Jenkins reports no attempt to standardize the interview questions across all of the selection boards.

Development of the test battery. Only two sources were located that provide any information on the development of the three tests comprising the Navy's pilot selection battery and its predictive validity (Fiske, 1947b; Liljencrantz, 1942). The information that could be gleaned from Liljencrantz's (1942) article is limited because it is based on a paper presented in early November 1942. Fiske (1947b) provides information on the predictive validity of the selection tests. The usefulness of this information is limited by the fact that the validities are for three cohorts accepted for flight training using different non-medical selection standards. Furthermore, the timeline is often lacking and some details pertaining to test development are vague. More importantly, Fiske does not fully identify the criteria used to assess predictive validity. Although "ground school" is a criterion, Fiske fails to identify which of the three ground school performances were predicted. Fiske does state that the predictive validity for flying was based on performance in "primary flight training." From Jenkins (1946) one can speculate that this nomenclature refers to Naval Air Station training.

According to Fiske (1947b), the initial pilot selection battery consisted of three tests, all of which were developed and/or validated with the help of the CSTAP. The first was the Wonderlic Personnel Test, a brief (12 min.) intelligence test. The biserial correlation for the least restricted cohort to pass/fail from ground school was r_{bis} = .31 and r_{bis} = .12 for flight failures with 2,356 students. The low predictive validity to flight failures combined with the demonstrated non-equivalence of the three versions of this test led the Navy to search for another test with better properties. Consequently, the Wonderlic was replaced in October, 1942 with the Aviation Classification Test, a longer (45 min.) intelligence test. Fiske presents no data on the predictive validity of this test.

The second test was a version of the Bennett Mechanical Comprehension Test created for the Navy by G.K. Bennett. This was a 45-min. test with minimal verbal content. Fiske (1947b) provides few details about the test characteristics except that the split-half reliability was .80 corrected for length. The test-retest (no interval given) reliability was between .84 and .87 (p. 602). The biserial correlation for the least restricted cohort was r_{bis} = .25 for ground school and r_{bis} = .33 for flight training. This test was retained through World War II.

The third test was a biographical inventory originally developed in 1939-1940 by the CSTAP for civilian pilots. A shorter version was developed for the Navy during 1940 and 1941 (Viteles, 1945). This initial version was administered on an experimental basis through 1941 (Fiske, 1947b) and some refinements may have occurred during this period. According to Fiske, the biographical inventory score was used in the decision to terminate or continue students who were doing poorly in flight training beginning in 1942. However, Liljencrantz indicates that scores from the biographical inventory and the intelligence test already were being used to eliminate students in 1941. Fiske (p. 611) shows biserial correlation for the least restricted cohort as r_{bis} = .06 for ground school and r_{bis} = .29 for flight training.

Beginning in December, 1942 scores from the biographical inventory were combined with those of the Mechanical Comprehension Test to produce a Flight Aptitude Rating (FAR) (Fiske, 1947b). The FAR score subsequently was used for selection into Navy pilot training. Fiske reports a predictive validity of r_{bis}= .43 between the FAR and pass/fail from flight training. Interestingly, Fiske makes no mention of the Aviation Classification Test being combined with the FAR although he does mention that scores on the Aviation Classification Test were used to reject applicants (p. 602). McFarland (1953) indicates that sometime later the Aviation Classification Test was added to the FAR, which increased the predictive validity of the FAR to r_{bis}= .50.

A striking difference between the pilot selection batteries of the AAF and the Navy concerns apparatus tests. One of the AAF Aviation Psychology Program Research Reports (Melton, 1947) is a 1,000+ page document devoted to the apparatus tests examined during the war. In contrast, few references mention apparatus testing for the Navy selection process. Clearly, the Navy became involved with apparatus tests before the war began. Jenkins (1941) mentions a CSTAP project for the Navy involving psychomotor tests. Viteles (1945) references a 1940 study conducted at the Naval Air Station at Pensacola, FL on the predictive validity of the Mashburn Serial Action Test, the Two-Hand Coordination Test, and an “eye-hand” coordination test.

McFarland (1953) states that the Serial Reaction Time Test, which may be the Mashburn Serial Action Test, increased the predictive validity of the FAR to r_{bis} = .61, but that the Navy rejected all psychomotor tests for practical, administrative reasons. However, Liljencrantz (1941) mentions that psychomotor tests were under investigation for track selection, that is, for assigning a student to bomber, fighter, or patrol aircraft.

Summary

During World War II, the Army and the Navy adopted different approaches to research on pilot selection and developed different types of selection batteries. The AAF employed a two-phase selection process. The first phase consisted of a basic physical examination and one written test, the AAFQE. Applicants that passed the first phase, became aviation cadets. The second phase consisted of a flight physical examination and the Aircrew Classification Battery. Initially, if candidates passed the physical examination, they were placed into one of three aircrew specialties: pilot, navigator, or bombardier based on their scores on the Aircrew Classification Battery and their preference. Applicants that failed the physical examination were assigned to a ground crew specialty. After August 1942 the Aircrew Classification Battery was also used to a limited extent as a selection device. By the end of the war, the number of aircrew specialties had been expanded to seven, including fighter pilot and bomber pilot.

Over the course of the war, the AAFQE was revised several times. The specific tests included in the Aircrew Classification Battery and their weights in the aircrew composites also were modified over time to increase the predictive validity of the battery. Individual selection instruments were frequently revised to increase their predictive validity. All of these revisions were based on extensive research conducted by uniformed psychologists.

Little documentation was found on the Navy's selection efforts. The Navy, like the Army, began working with the CSTAP in 1939. Unlike the Army, the Navy appears to have relied on the CSTAP throughout much of the war for selection research. The few articles that could be located describing the Navy's pilot selection effort show they had fewer uniformed psychologists than the Army. More importantly, only a small number of these psychologists were concerned directly with pilot selection; many were assigned to test administration and to selection of other aircrew members, such as gunners. The non-medical portion of the Navy pilot selection battery consisted of three tests: measures of general intelligence, mechanical aptitude, and a biographical data. A striking difference between the AAF's and the Navy's selection and classification batteries concerned apparatus tests. The AAF always had at least five apparatus tests in its classification battery, the Navy had none.

By the end of World War II, pilot selection had been placed on a strong scientific foundation. In contrast to the research conducted in World War I, the studies conducted in World War II used "modern" experimental and statistical methods, which assured the robustness of the results and the continued usefulness of the data. Perhaps the greatest contribution of the pilot selection effort was to the professional development of numerous psychologists who made subsequent major contributions to the areas of intelligence, statistical methods, and human factors.

BETWEEN THE WARS II

The journal articles concerned with pilot selection published in the years immediately following World War I described activities conducted during the war. This type of publication lag following a war is to be anticipated; wartime activities must be documented and information, declassified. Additionally, even in the early 1920's, journals had a publication lag. Thus, the articles pertaining to the World War I pilot selection efforts continued to be published for approximately 4 years after the end of the war.

One could anticipate the same publication pattern following World War II and, in fact, the pattern does mirror that of World War I closely. The major difference lies in the fact that 22 years separated World I and II, whereas only 5 years separated World War II and the Korean War. Consequently, the majority of relevant documents published between 1945 and 1950 dealt with World War II efforts and are cited in the preceding chapter. Only one short article (Roff, 1948) dealt with new pilot selection research conducted between 1945 and the start of the Korean War on June 25, 1950. Roff documents a program of basic research designed to identify the abilities that underlie performance on the pilot selection and classification instruments. Some of the abilities that were being investigated were memory, spatial processing, visual perception, and reasoning. Interestingly, he also mentions efforts to modify the psychomotor devices to allow them to be distributed to testing centers around the United States.

Rogers, Roach, and Short(1986) provide a brief summary of the period from the end of World War II to 1950. From late 1945 until mid-1947, the number of pilot applicants decreased significantly. The Air Force, which was established as a separate service in 1947, adopted the same solution to the decreasing pool as the Navy and the Army did in the late 1930's: traveling selection boards. However, the post World War II pilot selection system included apparatus tests that had to be transported to the testing locations. Apparatus-based testing was discontinued in 1955 because it was difficult to standardize administration procedures and to calibrate the electro-mechanical equipment used to administer the tests under decentralized testing conditions (Passey & McLaurin, 1966). In 1947 the applicant population had decreased to the point that the Air Force took any applicant who could pass the Air Force Qualifying Examination (AFQE, the revised version of the AAFQE) and had 2 years of college or the equivalent. Consequently, the Aircrew Classification Battery was discontinued.

MODERN PERIOD

Much research has been conducted since 1950 pertaining to pilot selection, and both the Air Force's and the Navy's pilot selection systems have undergone substantial changes as a result. It is beyond the scope of this effort to review all of this work. Instead, certain "themes" that have been important in pilot selection in the past and appear to be of importance today will be reviewed. These four themes are personality, biodata, timesharing, and identification of the personal attributes that lead to success as a pilot. To review these themes, representative Air Force and Navy reports will be discussed. More emphasis, however, will be placed on studies published in refereed journals. Literature reviews and meta-analyses will be especially emphasized.

Personality

Uniformed psychologists in World War I clearly were interested in personality as a predictor of flight success (e.g., Stratton et al., 1920). Personality assessments, however, were left to the physicians administering the physical examination; and, unlike biodata, no paper-and-pencil tests of personality were developed. In the early 1930's, De Foney (1931, 1933) demonstrated promising results between personality assessments conducted during the physical examination and subsequent accidents. This line of research appears not to have been continued. In World War II AAF psychologists evaluated personality by administering several commercially available tests, conducting interviews, and performing direct observations of candidates in stressful situations. None of these methods demonstrated usable predictive validities for success in flight training.

Dolgin and Gibb (1989) provide a review of the use of personality and interest measures from World War II to the mid-1980's. Those instruments concerned with personality generally show disappointing predictive validity. Dolgin and Gibb suggest that these failures may be attributed to three major factors. First, many of the personality assessments rest on subjective judgments, which are inherently unreliable. Second, many personality instruments are transparent.That is, the best response can be identified easily by candidates. Third, candidates for military flight training represent a select sample that presents restriction-in-range issues.

Siem (1992) administered an automated personality inventory to 509 undergraduate pilot trainees. The inventory consisted of 202 items from five different personality assessments. The items comprised 16 scales. A factor analysis of the items identified five major factors. Although all five of the factors had significant correlations with pass/fail from flight training, none provided significant incremental validity above that of the Air Force Officer Qualifying Test (AFOQT), a paper-and-pencil cognitive test which assesses verbal, math, spatial, aviation knowledge, and perceptual speed; and the Basic Attributes Test (BAT), an experimental battery assessing information processing and psychomotor coordination. A similar lack of incremental validity was demonstrated by Walters, Miller, and Ree (1993) using structured interviews on a comparable population.

Three meta-analyses have been conducted on personality since Dolgin and Gibb's (1989) review. The first, Hunter and Burke (1994), examined 68 studies published between 1940 and 1990,

some of which were conducted on civilian pilots. Hunter and Burke examined different categories of predictors, such as spatial ability and aviation information, as well as personality and biodata. The analysis included 46 validity coefficients and showed limited usefulness for personality measures with the confidence interval including 0.0 and $r_{mean} = .10$, the lowest of any category of predictor. The second, Martinussen (1996), analyzed data from 50 studies, all of which apparently used military personnel. Again, personality assessments showed the poorest predictive validities of any class of predictor (intelligence, psychomotor, etc.) with the lower confidence interval including 0.0 and $r_{mean} = .13$. Marinussen found that the predictive validity of personality measures was negatively correlated with the year of article publication but not significantly so.

Campbell, Castaneda, and Pulos (2010) criticized both the Hunter and Burke (1994) and the Martinussen (1996) analyses on the grounds that the personality assessments were aggregated into one effect, that is, the scales were not analyzed separately. Additionally, Hunter and Burke did not account for the fact that some scales should correlate positively with training outcome, whereas others should correlate negatively. Consequently, Campbell et al. disaggregated the personality scales and performed their meta-analysis on eight studies conducted only on military pilots. Because of the small number of studies included in the analysis, only extroversion, neuroticism, and anxiety (a facet of neuroticism) were analyzed. The criterion was pass/fail from flight training. Although all of the relations were in the hypothesized direction, the effect sizes were on the same order as those found by Hunter and Burke (1994) and Martinussen (1996), that is, barely different from 0. Campbell et al. suggest, like Dolgin and Gibb (1989), that restriction in range and item transparency may be responsible for the poor observed predictive validity.

The development of the Five Factor Model (e.g., McCrae & Costa, 1997), which provides a strong theoretical framework for understanding personality, has not resulted in an increase in the predictive validity of personality assessment for aviation selection (see Anesgart & Callister, 1999 for an exception). It remains to be seen if methodological improvements and more refined criteria will result in an acceptable incremental validity for personality measures.

Biodata

Biodata (biographical data) has had a long history of success in predicting pass/fail in flight training. Its usefulness was recognized first in World War I when athletic achievement was found to be related to success in flight training. A biographical inventory also was included in both the World War II AAF Aircrew Classification Battery and the Navy pilot selection battery. Both services continued to use a biographical inventory after the war.

In 1978 the biographical inventory was removed from the Air Force pilot selection battery because of an insufficient amount of female data for making selection decisions. The Navy also discontinued its use of its biographical inventory after the Vietnam War, but the exact date of the discontinuance could not be determined. North and Griffin (1977) show that the biographical inventory was still used to calculate the FAR and had an uncorrected predictive validity of $r = .19$ to pass/fail from flight training. Research by Street and Dolgin (1992) was designed to improve the current Navy biographical inventory. However, sometime after Street and Dolgin completed their study, the biographical inventory was eliminated from the Navy pilot selection system. No explanation for this elimination was found. It is interesting to note that the Alternate

Flight Aptitude Selection Test (AFAST; HQDA, 1987), the Army's pilot selection battery, still contains biographical items.

Youngling, Levine, Mocharnuk, and Weston (1977) reviewed studies examining biodata that were conducted from 1941 to 1974. Ten of the 13 studies reviewed were concerned with Air Force pilot selection; one, with Army selection; and the remaining two, with Navy selection. All of the studies examined the predictive validity of the biodata instruments to some measure of training performance. The biodata showed modest predictive validities (*rs* = .10 to .20) and had the most success predicting pass/fail from flight training.

Both Hunter and Burke (1994) and Martinussen (1996) included biodata measures in the meta-analyses described earlier. Hunter and Burke found that biodata predicted success in flight training (r_{mean} = .27). However, the predictive validity of biodata fell significantly from 1940 (r_{mean} = .30) to 1990 (r_{mean} = .09). These results were based on 21 validities. Martinussen's (1996) meta-analysis showed a mean r_{mean} = .21 based on 13 validities, exceeding that of both the personality and intelligence measures with r_{mean} = .13 for both. Like Hunter and Burke, Martinussen found a statistically significant negative correlation between the predictive validity and the year of article publication.

Interest in biodata as a non-cognitive selection instrument for the civilian sector has continued (Stokes, Mumford, & Owens, 1994). Nevertheless, a major drawback to the use of biodata instruments relates to scoring. The *Civil Rights Act of 1991* forbids the use of different cutoff points or adjustment of test scores based on sex. Any biodata items assessing athletic activities or interests must be carefully constructed to reduce sex differences in responses.

Timesharing

Theoretically, timesharing and multi-limb coordination are distinct constructs. In reality they are often difficult to separate. Many tests that assess timesharing require responses from two or more limbs. If these responses need to be coordinated in some fashion, the test may assess multi-limb coordination as well as timesharing. Similarly, tests that assess multi-limb coordination often have multiple stimuli, each of which requires a response from a different limb. Depending on the complexity of the responses and the temporal relation between the stimuli, the examinees may perform the task as if they were performing several, simpler tasks. An example of this is the Complex Coordination Test used in World War II. Consequently, it is important to remember that a timesharing test may also assess multi-limb coordination to some degree and vice versa.

No measures of timesharing ability appear to have been used in World War I. Mashburn may have assessed timesharing to a limited degree in his Serial Action Test, but this test does not appear to have been used for selection before World War II. In World War II, the Serial Action Test, now known as the Complex Coordination Test, was included in the Aircrew Classification Battery and may have assessed timesharing to a limited degree. The most direct measure of timesharing was the Rotary Pursuit Task with Divided Attention Test. As noted previously, this test had a predictive validity of r_{bis} = .29 for pass/fail from elementary flight training in the unselected airmen study and retained some predictive validity to more advanced stages of training.

Apparatus of some type is required to assess timesharing. Paper-and-pencil tests do not assess timesharing. After the Air Force removed apparatus tests from the pilot selection system, it could not assess timesharing. With the implementation of the BAT as an adjunct for plot training selection in 1993, timesharing was re-introduced into the Air Force's selection battery. The Test of Basic Aviation Skills (TBAS), which replaced the BAT in 2006, also includes tests of timesharing.

The incremental validity of timesharing measures to a selection system that assesses reaction time and information processing appears minimal. A meta-analysis conducted by Damos (1993) showed a statistically significant improvement in predictive validity for multiple-task measures as compared to single-task measures. However, the improvement, from $r = .18$ to $r = .23$ may be of little practical significance.

Few advances in the theoretical understanding of timesharing have occurred in the last 30 years. Advances may have been slowed by three problematic areas. First, the existence of a distinct timesharing ability is still questionable. The literature through 1990 shows equivocal evidence (See Brookings & Damos, 1991for a review). Subsequently, Carroll (1993) purportedly identified a timesharing factor. However, his identification was based on one dataset and is less than convincing. Second, the methodological problems associated with single- and multiple-task practice and scoring multiple-task performance have not been resolved (Damos, 1991). Third, statistical analysis problems associated with identifying a timesharing factor are particularly intractable (See Ackerman, Schneider, & Wickens, 1984).

Identification of Human Abilities

Arguably, one of the greatest contributions of the World War II pilot selection effort was an indirect one, the development of taxonomies of human abilities. J. P. Guilford was a director of one of the Psychological Research Units of the AAF in World War II and later authored *Printed Classification Tests* with J. I. Lacey. Based on some of the work conducted in World War II, Guilford (1967) developed a new theory of human intelligence that was seminal.

Fleishman developed his taxonomy of human abilities (Fleishman & Reilly, 2001) based in part on the work conducted by the Psychological Research Units in World War II. This taxonomy has had a profound impact on many areas of psychology and has been used to identify the abilities required for many tasks and jobs. Most importantly, it is the most commonly used taxonomy for job analyses of American civilian and military pilots. Damos (2011) provides recent military examples. Fleishman's list of abilities will be expanded shortly in the non-cognitive areas.

Perspective

One fact is evident in reviewing the development of the initial selection battery for both World War I and II: The investigators included tests in a battery because they had the technology available to do so, not because there was an apparent need. In World War I this approach was understandable. In World War II, this approach was less justifiable given the emphasis on performing job/task analysis and identifying the cause of failures in flight training. Although this report does not cover the period from 1950 to the present in detail, the persistent efforts

devoted to assessing personality traits, background information, and timesharing ability points toward a continued research agenda other than an understanding of the job.

Damos (2011) reviewed all available job analyses for fixed- and rotary-wing aircraft. Only nine studies were located. The earliest was conducted in 1960 and the most recent, in 2009, a span of almost 50 years. For a variety of reason, several of these had serious methodological shortcomings that limit their usefulness. The small number of studies points toward a need to conduct comprehensive job analyses for fixed-wing aircraft that identify the required knowledge, skills, abilities and other characteristics (KSAOs) and can guide research on pilot selection. Howse (2011), in a similar review of job analyses for unmanned aerial vehicle (UAV) operators, located only eight relevant studies and found similar methodological shortcomings. Again, a comprehensive job analysis is needed.

Why have sufficiently comprehensive job analysis not been performed? Several reasons can be given. First, although all military aircraft appear to require a core of KSAOs—such as multi-limb coordination, flying proficiency, and knowledge of communication procedures—different aircraft may require some KSAOs that are not common to all aircraft. Additionally, some aircraft may require more of a specific ability or skill than other aircraft. For example, helicopters may require more multi-limb coordination than large transport aircraft. A single job analysis for use in developing a selection system for different categories of aircraft (fighter, transport, etc.) may not be sufficient. Second, Fleishman's taxonomy has been used for many years and provides the basis for most job analyses conducted for pilots, e.g., Houston and Bruskiewics (2006). However, the cognitive ability section of his taxonomy is probably incomplete; Carroll (1993), for example, presents evidence for many more cognitive abilities than Fleishman includes in his taxonomy. Third, the increase in cockpit automation may require a different methodological approach to job analysis. For example, cognitive task analysis may become a more important component of a job analysis for a pilot.

If substantial progress is to be made in pilot selection, an updated taxonomy of human abilities must be constructed. This may require more basic research on the structure of human abilities. It may also require confirmation of the abilities identified by investigators such as Carroll (1993). Methodological improvements to job analysis also are needed to deal more effectively with the cognitive aspects of flying. Although many methodological advances have been made in cognitive task analysis, further work is necessary to develop cognitive task analysis as a generic tool.

REFERENCES

Ackerman, P. L., Schneider, W., & Wickens, C. D. (1984). Deciding the existence of a time-sharing ability; A combined methodological and theorectical approach. *Human Factors, 26*, 71-82.

Anderson, H. G. (1919). *The Medical and Surgical Aspects of Aviation*. London: Henry Frowde Hodder & Stoughton.

Anesgart, M., & Callister, J. D. (1999). Predicting training success with the NEO: The use of logistic regression to determine the odds of completing a pilot's screening program *Proceedings of the Tenth International Symposium on Aviation Psychology*. The Ohio State University, Columbus.

Annual Report of the Secretary of War for the Year 1885. Vol III: Report of the Chief of Ordnance. (1885). Washington, DC: U.S. Govenment Printing Office.

Anon. (1919). *Air Service Medical*. Washington, D.C.: War Department: Air Service. Division of Military Aeronautics Government Printing Office.

Armstrong, H. G. (1943). *Principles and Practice of Aviation Medicine* (Second ed.). Baltimore, MD: The Williams & Wilkins Company.

Bigelow, R. B. (1940). The evaluation of aptitude for flight training: The Rorschach method as a possible aid. *Journal of Aviation Medicine, 11*, 202-209.

Brookings, J. B., & Damos, D. L. (1991). Individual differences in multiple-task performance. In D. L. Damos (Ed.), *Multiple-Task Performance* (pp. 363-386). London: Taylor & Francis.

Campbell, J. S., Castaneda, M., & Pulos, S. (2010). Meta-analysis of personality assessments as predictors of military aviation training success. *International Journal of Applied Aviation Studies, 20*(1), 92-109.

Carlson, W. A. (1939). A proposed aid in the selection of Army flying cadets. *Journal of Aviation Medicine, 10*, 66-71.

Carlson, W. A. (1941). Intelligence testing of flying cadet applicants: A report on psychometric measurement. *Journal of Aviation Medicine, 12*(3), 226-229.

Carroll, J. B. (1993). *Human cognitive abilities A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Damos, D. L. (1991). Dual-task methodology: Some common problems. In D. L. Damos (Ed.), *Multiple-Task Performance* (pp. 101-119). London: Taylor & Francis.

Damos, D. L. (1993). Using meta analysis to examine the predictive validity of single- and multiple-task measures to flight performance. *Human Factors, 35*, 615-628.

Damos, D. L. (2011). *KSAOs for Military Pilot Selection: A review of the Literature*. Randolph AFB, Texas: Air Force Personnel Center Strategic Research and Assessment.

Davies, W. W. (1940). Some observations on aviation cadet selection. *Journal of Aviation Medicine, 11*, 37-42.

Davis, F. B. (1947). *The AAF qualifying examination. Report No. 6*. Washington, D. C.: U.S. Government Printing Office.

De Foney, C. G. (1931). A psychological study made on candidates for aviation training. *U.S. Naval Medical Bulletin, 29*(2), 191-204.

De Foney, C. G. (1933). A second psychological study made on candidates for aviation training. *U.S Naval Medical Bulletin, 31*(2), 103-111.

Deemer, W. L. (1947). *Records, analysis, and test procedures. Report No. 18*. Washington, DC: U.S. Government Printing Office.
Deemer, W. L., & Rafferty, J. A. (1948). Experimental evaluation of the psychiatric interview for prediction of success in pilot training. *Journal of Aviation Medicine, 20*, 238-250.
Dockeray, F. C. (1920). Department of Psychology *Aviation Medicine in the A.E.F.* (pp. 113-132). Washington: Government Printing Office.
Dockeray, F. C., & Isaacs, S. (1921). Psychological research in aviation in Italy, France, England and the American Expeditionary Forces. *Comparitive Psychology Journal, 1*(2), 115-148.
Dolgin, D. L., & Gibb, G. D. (1989). Personality assessment in aviator selection. In R. S. Jensen (Ed.), *Aviation psychology.* (pp. 288-320): Gower Publishing Co.
DuBois, P. H. (Ed.). (1947). *The Classification Program* (Vol. 2). Washington, D.C.: U.S. Government Printing Office.
Dunlap, K. (1917). Association-reaction as a test of learning. *Journal of Experimental Psychology, 2*(5), 386-391.
Farr, W. D. (1993). For want of a flight surgeon... *Aviation, Space and Environmental Medicine, 64*(5), 405-408.
Fiske, D. W. (1946). Naval aviation psychology, III: The special services group. *American Psychologist, 1*, 544-548.
Fiske, D. W. (1947). Naval aviation psychology, IV: The central research groups. *American Psychologist, 2*, 67-72.
Fiske, D. W. (1947). Validation of Naval Aviation Cadet Selection Tests against training criteria. *Journal of Applied Psychology, 31*, 601-614.
Flanagan, J. C. (1942). The selection and classification program for aviation cadets (aircrew-bombardiers, pilots, and navigators). *Journal of Consulting Psychology, 6*, 229-240.
Flanagan, J. C. (1947). *The AAF Qualifying Examination Report No. 6* (Vol. 6): U.S. Government Printing Office, Washingtion 25, D.C.
Flanagan, J. C. (1948). *The Aviation Psychology Program in the Army Air Forces Report No. 1* (Vol. 1). Washington, D.C.: U.S. Government Printing Office.
Fleishman, E. A., & Reilly, M. E. (2001). *Handbook of Human Abilities*. Potomac, MD: Management Research Institute, Inc.
Glenn, C. R. (1935). A preliminary report on a performance test for flying. *Journal of Aviation Medicine, 6*, 14-19.
Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw Hill.
Guilford, J. P., & Lacey, J. I. (Eds.). (1947). *Printed classification tests. Report No. 5*. Washington, D.C.: U.S. Government Printing Office.
Henmon, V. A. C. (1919). Air service tests of aptitude for flying. *Journal of Applied Psychology, III*(2), 103-109.
Houston, J. S., & Bruskiewics, K. T. (2006). *Development and preliminary validation of a selection instrument for U.S. Army flight training (SIFT): Volume 2*.
Howse, W. R. (2011). *KSAOs for Unmanned Aircraft Operators* (No. DAS-2011-04). Gurnee, IL: Damos Aviation Services, Inc.
HQDA. (1987). *Alternate Flight Aptitude Selection Test (AFAST) Information Pamphlet*.
Hunter, D. R., & Burke, E. F. (1994). Predicting Aircraft Pilot-Training Success: A Meta-Analysis of Published Research. *International Journal of Aviation Psychology, 4*(4), 297.
Jenkins, J. G. (1941). Selection and training of aircraft pilots. *Journal of Consulting Psychology*, 228-234.

Jenkins, J. G. (1945). Naval aviation psychology, I: The field service organization. *Psychological Bulletin, 42*, 631-637.
Jenkins, J. G. (1946). Naval aviation psychology, II: The procurement and selection organization. *American Psychologist, 1*, 45-49.
Johnson, H. M. (1920). Resume of research in the psychology of aviation during the year 1919. *Science, LI*(1323), 449-452.
Johnson, R. H. (1917). Select Army aviators by test, not education. *Journal of Heredity, 8*, 425.
Jones, D. R. (2008). Flying and Dying in WWI: British Aircrew Losses and the Origins of the U.S. Military Aviation Medicine. *Aviation, Space and Environmental Medicine, 79*(2), 139-146.
Kaufman, B. (1943). Notes on classification, selection and training. *Journal of Aviation Medicine, 14*(6), 383-385.
Liljencrantz, E. (1942). Problems in the selection of aviators. *Journal of Aviation Medicine, 13*, 107-120.
Martinussen. (1996). Psychological measures as predictors of pilot performance: A meta analysis. *International Journal of Aviation Psychology, 6*, 1-20.
Mashburn, N. C. (1934a). The complex coordinator as a performance test in the selection of military flying personnel. *Journal of Aviation Medicine, 5*, 145-154.
Mashburn, N. C. (1934b). Mashburn automatic serial action apparatus for detecting flying aptitude. *Journal of Aviation Medicine, 5*, 155-160.
Mashburn, N. C. (1935). Some interesting psychological factors in the selection of military aviators. . *Journal of Aviation Medicine, 6*, 113-126.
Mashburn, N. C. (1939). The selection of the trainee for military aviation. *The Military Surgeon, 84*, 428-441.
McComas, H. C. (1922). *The aviator*. New York: E.P. Dutton & Company.
McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 57*(5), 509-516.
McFarland, R. A. (1953). *Human Factors in Air Transportation*. New York: McGraw Hill.
Melton, A. W. (1943). The selection of pilots by means of psychometric tests. *Journal of Aviation Medicine, 15*, 116-123.
Melton, A. W. (Ed.). (1947). *Apparatus tests. Report No. 4*. Washington, DC: U.S. Government Printing Office.
Norman, R. D. (1947). Comparison of earlier and later success in Naval aviation training. *Journal of Applied Psychology, 31*, 511-518.
North, R. A., & Griffin, G. R. (1977). *Aviator Selection 1919 - 1977* (No. Special Report -77-2). Pensacola, FL: Naval Aerospace Medical Research Laboratory.
Office of the Air Surgeon. (1944). The aviation cadet qualifying examination of the Army Air Forces. *Psychological Bulletin, 41*(6), 385-394.
Ovington, E. (1914). The Psychic Factors in Aviation. *Journal of the Americal Medical Association, LXIII*(5), 419-420.
Passey, G. E., & McLaurin, W. A. (1966). *Perceptual-psychomotor tests in aircrew selection: Historical review and advanced concepts* ((PRL-TR-66-4). Lackland AFB, TX: Personnel Research Laboratory.
Poppen, J. R. (1941). Recent trends in aviation medicine. *Journal of Aviation Medicine, 12*, 53-71.
Razran, G. H. S., & Brown, H. C. (1941). Aviation. *Psychological Bulletin, 38*, 322-330.

Roff, M. F. (1948). Psychological research at the AAF School of Aviation Medicine. *Journal of Aviation Medicine, 19*, 20-23.

Rogers, D. L., Roach, B. W., & Short, L. O. (1986). *Mental ability testing in the selection of Air Force Officers: A brief historical overview* (Technical Paper No. AFRL-TP-86-23). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Siem, F. M. (1992). Predictive Validity of an Automated Personality Inventory for Air Force Pilot Selection. *International Journal of Aviation Psychology, 2*(4), 261.

Staff of the Psychological Section. (1945). Psychological activities in the Training Command, Army Air Forces. *Psychological Bulletin, 42*, 37-54.

Staff Psychological Reseach Unit #1. (1945). History, organization, and research activities, psychological research project. *Psychological Bulletin, 42*, 751-759.

Stokes, G. S., Mumford, M. D., & Owens, W. A. (Eds.). (1994). *Biodata Handbook*. Palo Alto, CA: CPP Books.

Stratton, G. M., McComas, H. C., Coover, J. E., & Bagby, E. (1920). Psychological tests for selecting aviators. *Journal of Experimental Psychology, 3*(6), 405-423.

Street, D. R., & Dolgin, D. L. (1992). *The efficacy of biographical inventory data in predicting early attrition in Naval aviation officer candidate training* (No. NAMRL-1373). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Sutton, D. G. (1930). Psychology in Aviation. *US Navy Medical Bulletin, 28*, 5-13.

Thorndike, E. L. (1919). Scientific personnel work in the army. *Science, 49*, 53-63.

Viteles, M. S. (1945). The Aircraft Pilot: 5 Years of Research A Summary of Outcomes. *Psychological Bulletin, 42*(8), 489-526.

Walters, L. C., Miller, M. R., & Ree, M. J. (1993). Structured Interviews for Pilot Selection: No Incremental Validity. *International Journal of Aviation Psychology, 3*(1), 25.

Wilmer, W. H., & Ireland, M. W. (1920). *Aviation Medicine in the A.E.F.* Washington, D.C.: War Department, Director of Air Service.

Yerkes, R. M. (1918). Psychology in relation to the war. *Psychological Review, 25*(2), 85-115.

Yerkes, R. M. (1919). Report of the Psychology Committee of the National Research Council. *Psychological Review, 26*, 83-149.

Youngling, E. W., Levine, S. H., Mocharnuk, J. B., & Weston, L. M. (1977). *Feasibility study to predict combat effectiveness for selected military roles: Fighter pilot effectiveness*. East St. Lous, MO: McDonnell Douglas (MDC E1634).

APPENDIX A

European Selection for World War I

Although this report is concerned with pilot selection in the United States, some of the US tests were obtained from our World War I Allies, specifically from the French and the Italians. Because the Allies entered the war in 1914, they had a 2- to 3-year head start on the Americans developing a pilot selection system. Fortunately for the US, this allowed the Allies to provide valuable information on the types of selection tests they found effective.

The Italians were the first to study the skills and abilities needed for success as a pilot (Dockeray & Isaacs, 1921). Based on a series of preliminary studies of good, average, and poor pilots, they determined that pilots needed 1) speed of perception, 2) distribution of attention, 3) coordinated psychomotor activity, and 4) a low level of emotional reactivity. Their studies led to the development of simple reaction time tests to auditory and visual stimuli. Norms subsequently were established. Candidates whose reaction times were too slow (greater than 0.2 s for visual stimuli and 0.17 s for auditory stimuli) or too variable were eliminated. These tests eliminated 247 of 13,936 (1.8%) candidates tested in 1918. Eventually, some of the Italian scientists argued that the cut off points should be based on the distribution of scores rather than on arbitrary values. However, the change to a distribution-based cutoff does not seem to have been adopted.

The Italians also developed several complex reaction time tests. One of these tests was a four-choice reaction time test. Two of the alternatives required manual responses and two, foot responses. Occasionally, the candidate had to respond to two stimuli simultaneously. Two versions of another test required the candidate to move a lever like a control stick in an aircraft in one of four directions (left, right, forward, backwards) in response to a visual stimulus. In one version of the test, a fifth stimulus indicated that the candidate was to make no response. A candidate's mean reaction time and variability were calculated, but no cut offs were reported.

Because emotional lability was of interest to the Italians, they developed a test that became known as the "surprise" test. This test began by having the test administrator collect breathing rate, pulse, and hand steadiness under normal (resting) conditions. After sufficient baseline data had been collected, the experimenter discharged a gun or a firecracker or played an automobile claxon behind the candidate and out of his sight. The candidate's breathing rate, vasomotor constriction, and startle response were recorded. No criteria were reported for eliminating candidates based on this method, but 232 candidates out of 13,936 (1.7%) were eliminated for excessive lability (Dockeray & Isaacs, 1920).

Later the Italians administered the "surprise" test between two sets of simple reaction time measures. The average score on the post-test series was compared to the pretest series. An increase of 25% or more in the average reaction time of the post-test series was disqualifying, but the reason for this cut off is not given (Dockeray & Isaacs, 1920).

Two other categories of tests were used. The first category consisted of one test, a cancellation test. For this the candidate received a sheet with irregularly placed symbols. The candidate had to mark certain symbols. The test was to be completed in 5 min with a maximum of five errors. Again, no rationale for this cutoff is given nor are any validity data presented. The second

category consisted of tests of perceptual speed. Two tests gradually increased the presentation time of visual stimuli until the candidate could identify the stimuli. Another test presented a series of simple, colored forms. The stimuli were exposed for 1 s and the candidate had an additional 1.5 s to identify the shape and color of the stimuli. The cut score was based on the number of errors in a series of 20 stimuli.

The French pilot selection system appears to have been developed by three prominent French psychologists--Camus, Nepper, and Binet. Like the Italians, they used tests of simple visual and auditory reaction time and may have obtained the tests from the Italians. They also included a test of simple reaction time to tactile stimuli. All reaction times were measured to millisecond accuracy and mean performance apparently was calculated on the basis of ten responses. Norms for the three modes of simple reaction time were based on experienced pilots. Dockeray (1920) reports that a candidate whose average reaction time was 100 ms longer than the mean of the experienced pilots was disqualified. If the candidate had one reaction time that was 100 ms or greater than his own average, he also was disqualified.

The French also used the surprise test. Anderson (1919) indicates that the time to return to baseline was the primary consideration; candidates who recovered rapidly were favored. He also implies that this test was used more for track selection rather than for primary selection; those candidates who recovered most quickly were assigned to pursuit aircraft. In contrast, McComas (1922) states that candidates with the smallest difference between the resting and the post-noise value on each dependent measure were assumed to have the highest likelihood of becoming good pilots. Thus, according to McComas the surprise test was a primary selection test, not a track selection test. Sometime between early 1917 and 1921, the surprise test was eliminated from the French pilot selection battery.

The British did not begin systematic selection of pilots until 1916 when the Medical Selection Board was established (McComas, 1922, p. 183). Anderson (1919), a British air surgeon, describes a visit to see Camus, Nepper, and Binet in early 1917. Later, he implies that the British did not adopt the reaction time tests developed by the French although he personally was favorably impressed. No evidence was found indicating that the US obtained any tests from the British.